Heavy Traffic Analysis of Dynamic Cyclic Policies: A Unified Treatment of theSingle Machine Scheduling Problem

David M. Markowitz and Lawrence M. Wein

#3925-96-MSA

September, 1996

HEAVY TRAFFIC ANALYSIS OF DYNAMIC CYCLIC POLICIES: A UNIFIED TREATMENT OF THE SINGLE MACHINE SCHEDULING PROBLEM

David M. Markowitz
Logistics Management Institute
McLean, VA 22102

Lawrence M. Wein
Sloan School of Management, M.I.T.
Cambridge, MA 02139

ABSTRACT

This paper examines how setups, due-dates and the mix of standardized and customized products affect the scheduling of a single machine operating in a dynamic and stochastic environment. We restrict ourselves to the class of dynamic cyclic policies, where the machine busy/idle policy and lot-sizing decisions are controlled in a dynamic fashion, but different products must be produced in a fixed sequence. As in earlier work, we conjecture that an averaging principle holds for this queueing system in the heavy traffic limit, and optimize over the class of dynamic cyclic policies. The results allow for a detailed discussion of the interactions between the due-date, setup and product mix facets of the problem.

September 1996

This paper focuses on what we consider to be the three most important structural features of single machine scheduling problems. The first characteristic is the presence or absence of setup costs and/or setup times when the machine switches from one type of product to another. Setup penalties force the scheduler to exploit the economies of scale, which leads to dynamic lot-sizing policies. The second factor is the presence or absence of advanced information regarding future demands, which gives rise to problems with and without due-dates, respectively. This aspect of the problem essentially dictates the nature of the objective function: If advanced information is provided then the objective function is based on due-date considerations (e.g., earliness and tardiness costs) and if no such information is available then the objective function is expressed in terms of system measures (e.g., inventory costs, waiting time, throughput). The third characteristic is whether products are customized or standardized. This feature is intimately related to the make-to-stock/make-to-order distinction: Customized products must be made-to-order whereas standardized products can (but do not need to) be made-to-stock.

The aim of this paper is to provide a unified (with respect to these three features) treatment of single machine scheduling in a dynamic and stochastic environment. More specifically, we consider a manufacturing system consisting of one machine that produces multiple types of goods, which we refer to as *products*. Each product can be either customized, which requires the request for an order before production can begin, or standardized, in which case they can be pre-stocked in a finished goods inventory. The machine is limited in capacity and can only produce one unit at a time. Whenever the machine switches from producing one product to another, a setup cost and/or setup time penalty is incurred. Orders arrive to the system, each requesting a unit of the product at a specified due-date. A completed unit assigned to an order before the order's due-date must be held and an earliness cost is incurred. Similarly, a tardiness fee is levied whenever a completed unit is delivered to an order after its due-date. In addition, unassigned items held in finished goods inventory also incur a holding cost equal to the earliness fee of that (standardized) product. Interarrival times, service times, due-date lead times (an order's due-date minus its time of arrival) and setup times are random variables that can vary by product.

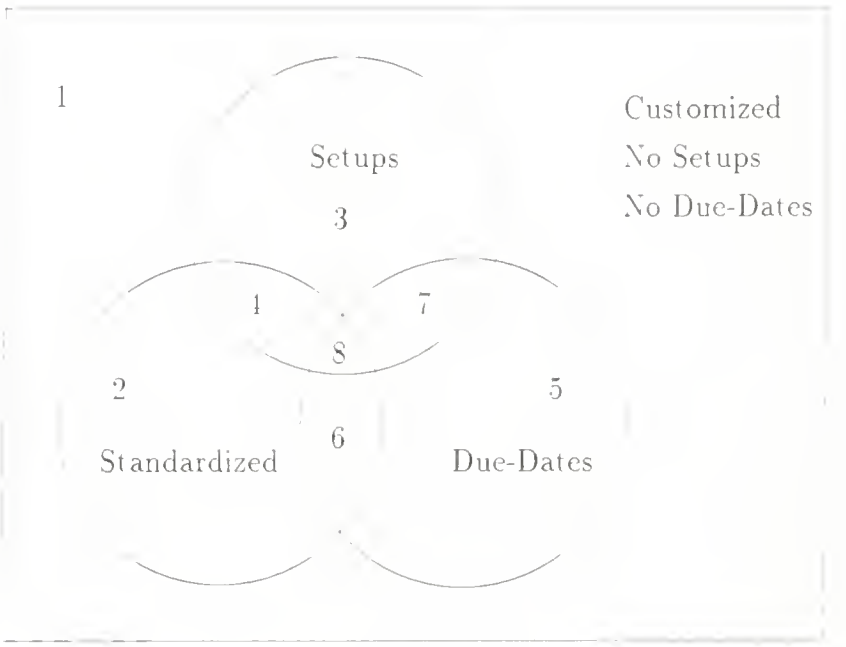

Figure 1: The eight subproblems under consideration.

In this setting, the machine at any point in time can idle, produce the product that it is currently set up for, or initiate a setup for a different product. In other words, the scheduler makes three types of decisions in a dynamic fashion: Whether the machine should be busy or idle, how much to make of each product (i.e., lot-sizing) and which product to produce next. In this paper, we restrict ourselves to *dynamic cyclic policies*, which allows full discretion over the first two decisions, but produces each product in a fixed cyclic sequence. We wish to optimize the queueing system with respect to long run expected average costs due to earliness, tardiness, holding and setups.

The Venn diagram in Figure 1 delineates the eight subproblems that are incorporated by our analysis. Although only the general scheduling problem is considered throughout the body of the paper, §5 is devoted to a discussion of each of the eight regions in the Venn diagram. Since the existing literature has only examined specific regions of the Venn diagram, we delay our literature review to this discussion.

This paper expands upon the methods of Markowitz, Reiman and Wein (1995) (abbreviated hereafter by MRW), which analyzes the stochastic economic lot scheduling

problem (depicted as subproblem 4 in Figure 1). MRW applies Coffman, Puhalskii and Reiman's (1995a, 1995b) *heavy traffic averaging principle*, which considers two sets of scalings: A fast one where time is sped up by a factor of $O(n)$ and a slow one where time is increased by a factor of $O(\sqrt{n})$ (where n goes to infinity in the heavy traffic limit). The fast scaling leads to a diffusion limit and the slow scaling leads to a fluid limit. The heavy traffic averaging principle couples these two processes and makes this difficult scheduling problem amenable to analysis. As in MRW, we optimize the control problem, first in the fluid limit and then in the diffusion limit. The primary analytical contribution of this paper is to determine how due-dates affect the system behavior under the fluid limit. As in our previous applications of the heavy traffic averaging principle (Reiman and Wein 1994, 1995, MRW and Reiman, Rubio and Wein 1995), we conjecture that this principle holds for the system under study, without providing a rigorous proof of convergence; see Reiman and Wein (1996) for a heuristic justification of this conjecture. Because a closed form solution has eluded us, we resort to developing a computational procedure for solving the control problem. However, to gain further insight, we analyze the special case where each product has a different deterministic (as opposed to random) due-date lead time; in this case, the results simplify considerably. Finally, a simulation study is performed to assess both the effectiveness of our proposed policies and the accuracy of the heavy traffic approximation.

The scheduling problem is formulated in §1 and heavy traffic preliminaries are introduced in §2. We perform a heavy traffic analysis of the scheduling problem in §3 and work through the deterministic due-date lead time case in §4. Section 5 is devoted to a discussion of the subproblems outlined in the Venn diagram in Figure 1 and §6 contains results of a computational study on due-date effects. Concluding remarks are offered in §7. Readers interested only in the qualitative insights derived from this work may omit §2, §3 and §4.

1. PROBLEM FORMULATION

A single machine produces N different products. Without loss of generality, we assume that products $1, 2, \ldots, N^c$ are customized and $N^c + 1, \ldots, N^c + N^s = N$ are

standardized. Each product i has its own generally distributed service time with mean μ_i^{-1} and coefficient of variation c_{is}. *Orders* for goods arrive from an exogenous renewal demand process. For each product i, the interarrival times of orders are generally distributed with mean λ_i^{-1} and coefficient of variation c_{ia}. The arrival and service time processes for each product are assumed to be independent, although they need not be (see Reiman 1984 for the incorporation of correlated compound renewal processes). The utilization of product i is $\rho_i = \lambda_i/\mu_i$ and the system utilization, or traffic intensity, is $\rho = \sum_{i=1}^{N} \rho_i$.

Orders arrive to the system with a specified due-date. The *due-date lead time*, which is the time interval between an order's arrival and its due-date, for product i is a random variable with density $\tilde{f}_i(s)$, cumulative distribution function $\tilde{F}_i(s)$ and mean $\tilde{l}_i$, and is independent of the arrival and service processes. (For quantities that undergo scalings, a "tilde" denotes a lack of scaling, a "bar" denotes the fluid scaling, and no symbol above the quantity denotes the diffusion scaling.) We assume that the due-date lead time distribution has bounded support, and denote its minimum and maximum by $\tilde{a}_i$ and $\tilde{b}_i$, respectively; because due-date information typically reflects future demand, we further assume that $\tilde{a}_i \geq 0$.

Orders for customized goods are immediately queued. The machine can only work on a customized product if an order is present. Once a customized order is serviced, the completed unit is either held until the order's due-date or delivered immediately if the order is tardy. If the unit is held, the order incurs an earliness fee (holding cost) $\hat{h}_i$ per unit time until the due-date; if the order is past due, it incurs a tardiness fee (backorder cost) of $\hat{b}_i$ per unit time late. In anticipation of our subsequent analysis, we also express these cost parameters in terms of units of work: $h_i = \hat{h}_i\mu_i$ and $b_i = \hat{b}_i\mu_i$.

The flow of physical product and orders is more complex for standardized products, and we need to keep track of both orders and completed *items*. Completed items can be pre-stocked in the finished goods inventory, where each unit accrues holding costs at a rate of $\hat{h}_i$ per unit time. Orders for standardized products are queued upon arrival, and can be filled by an item either from the finished goods inventory or directly from the output of the machine. Once an item is assigned to an order, the system incurs either an earliness

fee $\hat{h}_i$ for each unit of time until the order's due-date or a tardiness fee of $\hat{b}_i$ per unit time tardy. Again, these costs have workload equivalents $h_i = \hat{h}_i\mu_i$ and $b_i = \hat{b}_i\mu_i$. Notice that we have made the natural assumption that the holding cost for an unallocated item in the finished goods inventory is identical to the earliness cost incurred when a completed item is allocated to a standardized order before its due-date (in the latter case, one can envision the item sitting on the shipping dock until the order's due-date). Hence, no benefit can be gained by assigning a completed item to a standardized order *before* its due-date: Delaying the allocation of completed items to orders results in more flexibility and a lower cost policy. Therefore, without loss of generality, we only consider policies that incur no earliness fees for standardized products, and so standardized orders exit the system when completed items are assigned to them.

The scheduler can observe the system state at each point in time. To ease the notational burden, we only introduce notation in this section for those quantities that will be used in our subsequent heavy traffic analysis. Let the *slack* of an order be its due-date minus the current time. Hence, an order's slack is identical to its due-date lead time when it arrives to the system, but the slack, which can be positive or negative, is a dynamic quantity that decreases at unit rate throughout the order's sojourn in the system. The system state includes the arrival time and the slack of each order in each product's queue, the number of items of each standardized product in finished goods inventory, which we denote by $\tilde{I}_i(t)$ for $i = N^c + 1, \ldots, N$, the product that is currently set up (or being set up), and the residual service and setup times (if they are currently in progress).

The machine follows a dynamic cyclic policy. All products are serviced in a fixed cycle. At any point in time the scheduler can deploy the machine in one of three ways: Produce the product currently set up for (this might not be possible if set up for a customized product and there are no orders present), set up for the next product in the cycle, or idle. The policy is dynamic in that the lot sizes and the busy/idle policy can be molded to address the changing needs of the system.

A penalty is incurred every time the machine switches production to the next product in the cycle. This penalty can be a cost, a period of downtime or both. The heavy traffic

performance of the system depends on these penalties only through the average setup cost per cycle, $\tilde{K}$, and the average total switchover time per cycle, s. If only one form of setup penalty is present and setups vary by product, then the best cycle order can be found by solving a traveling salesman problem, where the distance between two cities corresponds to the setup penalty between two products. If both forms of penalties are present then the situation is more complex. We also assume that the policy is preemptive-resume, but the approximation scheme used here is too coarse to differentiate between a system with a preemptive-resume policy versus one without preemption.

We wish to minimize the long run expected average cost of the system, and additional notation is required to describe our objective function. Let T_{ni} be the time that the n^{th} unit of product i is assigned to an order. When a completed unit is assigned to an order, we assume that it is assigned to the order with the earliest due-date within its product, as this will minimize cost (see Pandelis and Teneketzis 1995 and Righter 1996). For $t \geq 0$, define the *minimum slack* $\tilde{L}_i(t)$ to be the smallest slack among all orders in product i's queue at time t. Finally, let $\tilde{J}(t)$ be the cumulative number of cycle completions by time t. The long run average cost is

$$\limsup_{T\to\infty} \frac{1}{T}\left(\sum_{i=1}^{N^c}\sum_{\{n|T_{ni}<T\}}[\hat{h}_i\tilde{L}_i^+(T_{ni})+\hat{b}_i\tilde{L}_i^-(T_{ni})]\right. \tag{1}$$
$$\left.+\sum_{i=N^c+1}^{N}[\int_0^T \hat{h}_i\tilde{I}_i(t)dt+\sum_{\{n|T_{ni}<T\}}\hat{b}_i\tilde{L}_i^-(T_{ni})]+K\tilde{J}(T)\right).$$

where $x^+(t)=\max\{x(t),0\}$ and $x^-(t)=\max\{-x(t),0\}$. Notice that the multiplicative product of cost and queue length is integrated over time for the items in finished goods inventory, whereas we have chosen to sum the product of cost and time (e.g. $\hat{h}_i\tilde{L}_i^+(T_{ni})$) over orders. This "reversal of the order of integration" is a necessary step in analyzing the effects of due-dates.

2. HEAVY TRAFFIC PRELIMINARIES

This section provides an overview of the heavy traffic analysis of the problem described in §1. The approach outlined here relies heavily upon results in Coffman, Puhal-

skii and Reiman and MRW, and we refer readers to these papers for more details.

Workload Processes. We begin by defining the key stochastic processes for our heavy traffic analysis. Recall that the system has queues of orders for all products and a finished goods inventory of items for all standardized products. For customized products $i = 1, \ldots, N^c$, let $\tilde{W}_i(t)$ be the total amount of time required for the machine to process all of the orders that are in product i's queue at time t. For standardized products $i = N^c + 1, \ldots, N$, let $\tilde{W}_i(t)$ be the total amount of time needed to process all standardized orders of product i *minus* the total machine time already invested in the units in product i's finished goods inventory at time t. We refer to $\tilde{W}_i = \{\tilde{W}_i(t), t \geq 0\}$ as the *workload process* for product $i = 1, \ldots, N$, and let $\tilde{W}^* = \sum_{i=1}^{N} \tilde{W}_i$ be the total workload process. The total workload at time t represents the total amount of work currently being requested (in the form of orders) minus the total work stored in inventory (this work will eventually be assigned to orders). This one-dimensional process is the natural definition of workload for systems with customized and standardized products; readers should note that, in contrast to MRW, the total workload is defined from the "make-to-order" point of view, in that work in orders is positive and work in finished goods inventory is negative.

Heavy Traffic Averaging Principle. Coffman, Puhalskii and Reiman's heavy traffic averaging principle (abbreviated by HTAP) has augmented the understanding of heavily loaded multiclass queueing systems that incur setup costs or setup times. Although rigorously proved for a two-class queue (in the absence and presence of setup times in 1995a and 1995b, respectively) that employs an exhaustive service discipline, numerical results in Reiman and Wein (1994, 1996), MRW and Reiman, Rubio and Wein support the conjecture that the HTAP holds for a much wider class of systems. As in these applications, we conjecture that the HTAP holds without providing a rigorous proof of convergence.

The HTAP is based on two sets of limits, taken as the total utilization goes to one and synchronized by a scaling parameter n. The first limit of the HTAP states that as $\sqrt{n}(1-\rho)$ approaches a constant c, the normalized total workload process $\tilde{W}^*(nt)/\sqrt{n}$ weakly converges to a diffusion process, $W^*(t)$, with parameters defined by the system data and

policy. It is called the *diffusion limit* and is the result of a functional central limit theorem. The second limit, called the *fluid limit*, states that for the same scaling parameter n and for a given total workload, the individual workloads $\tilde{W}_i(\sqrt{n}t)/\sqrt{n}$ converge almost surely to $\bar{W}_i(t)$, a fluid process; this result is related to the functional strong law of large numbers.

The *time scale decomposition* inherent in these two limits is intuitive: As utilization approaches one, the amount of work in the system is large and the total workload cannot change quickly; yet, as the machine switches between products, work can shift rapidly among the individual queues and inventories. The shifting of individual workloads occurs an order of magnitude, $O(\sqrt{n})$, more quickly than the total workload, and so the fluid limit evolves for a period while the total workload remains relatively constant. For example, the total workload might change on the order of weeks, while individual workloads change daily.

Dynamic Cyclic Policies and the Fluid Limit. Under a dynamic cyclic policy, a service *cycle* corresponds to the setting up and processing of each of the N products. If the machine is following a dynamic cyclic policy, then the HTAP implies that many cycles will be completed before there is a significant change in total workload $W(t)$. Because there are many cycles for a given total workload, a dynamic cyclic policy can be expressed as a function of only the total workload level, not the individual workloads; i.e., the lot sizes for each product and the busy/idle policy depend only on the total workload level. MRW show that dynamic cyclic policies in heavy traffic can be completely characterized by an idling threshold w_0 and two workload-dependent functions, the $N-$dimensional cycle center $x^c(w)$ and the cycle length $\tau(w)$ (throughout, we denote the total workload process by W and an arbitrary feasible total workload by w). The fluid process is unaffected by the idling threshold, and is dictated by the latter two functions. The i^{th} element $x_i^c(w)$ of the cycle center is the average amount of work in product i over the course of a cycle in the fluid limit. The cycle length $\tau(w)$ is the amount of time required to complete a cycle in the fluid limit. For the inventories and queues to remain balanced, the machine must allocate $\rho_i\tau(w)$ units of time per cycle producing product i. While product i is in service, the process $\bar{W}_i$ decreases at rate $(1-\rho_i)$; while the machine is set

up for other products. W_i increases at rate ρ_i (perhaps by depleting the finished goods inventory for $i = N^c+1, \ldots, N$). Setup times are unscaled (i.e., remain $O(1)$) and vanish in the fluid limit: The large amount of work in the system causes setups to be performed relatively infrequently. Thus, product i's workload level fluctuates by $\rho_i(1-\rho_i)\tau(w)$ over the course of a cycle, and ranges from a minimum of $x_i^c(w) - \rho_i(1-\rho_i)\tau(w)/2$ to a maximum of $x_i^c(w) + \rho_i(1-\rho_i)\tau(w)/2$. The cycle center and cycle length are parameters that can be set by the scheduler, subject to some restrictions: The sum of the cycle centers must equal the total workload, $\sum_{i=1}^{N} x_i^c(w) = w$; the minimum amount of work over the cycle must be nonnegative for a customized product, $x_i^c(w) - \rho_i(1-\rho_i)\tau(w)/2 \geq 0$; and if there are setup penalties then the cycle length $\tau(w)$ must be greater than zero.

Cost per Cycle. Given the dynamics described above, we can express the cost of a fluid cycle, $c(w)$, in terms of the parameters $x^c(w)$ and $\tau(w)$. The cost per cycle is composed of individual product (earliness, tardiness and holding) costs and setup costs, and is given by

$$c(w) = \sum_{i=1}^{N^c} c_i(x^c, \tau, w) + \sum_{i=N^c+1}^{N} c_i(x^c, \tau, w) + \frac{K}{\tau(w)}, \tag{2}$$

where $c_i(x^c, \tau, w)$ is product i's average cost rate for a cycle given a policy x^c and τ and total workload w, and $K = \tilde{K}/n$ is the normalized setup cost per cycle (see Reiman and Wein 1994 for a justification of this scaling). In MRW, we calculated $c_i(x^c, \tau, w)$ by integrating the cost over one production cycle and then dividing by the cycle length. In the next section, we perform the same type of calculation for our due-date problem but "interchange the order of integration" as discussed before. That is, instead of integrating over time in the cycle, we integrate over the orders filled during the cycle.

Dynamic Cyclic Policies and the Diffusion Limit. Coffman, Puhalskii and Reiman show that the drift of the diffusion process for total workload level w is $s/\tau(w) - c$. When there are no setup times, the diffusion limit is a reflected Brownian motion (RBM), regardless of the policy in use; this result coincides with classic results (Iglehart and Whitt 1970). Since the cycle length $\tau(w)$ is dictated by the dynamic cyclic policy, the drift of the diffusion process depends on both the policy and setup times when setup times are

present. The variance of the diffusion process is $\sigma^2 = \sum_{i=1}^{N} \frac{\lambda_i}{\mu_i^2}(c_{ai}^2 + c_{si}^2)$, and thus is only dependent on system parameters.

Although the idling threshold w_0 does not affect the fluid process, this control parameter does impact the one-dimensional diffusion process W by acting as its reflecting barrier. There are restrictions on w_0 only if there are no standardized goods in the system. In this case, w_0 must be nonnegative.

The Optimization Problem. Suppressing the notation illustrating the dependence of the cost on our policy, we can express the objective function in equation (1) as

$$\limsup_{T\to\infty} \frac{1}{T} \int_0^T c(W(t))dt, \tag{3}$$

where $c(w)$ is given in (2). Our controls are the idling threshold w_0 and the $N-$dimensional functions $x^c(w)$ and $\tau(w)$. As in MRW, the optimization is carried out as follows. We minimize (2) with respect to x^c to find the cycle center in terms of $\tau(w)$ and w_0. This is a constrained nonlinear optimization problem. Then (3) becomes a diffusion control problem with a drift control (via the cycle length $\tau(w)$) and a singular control (via the idling threshold w_0).

3. OPTIMAL DYNAMIC CYCLIC POLICIES

The goal of this section is to optimize the cost in (3). This optimization is performed in several steps: In §3.1 and §3.2, we calculate the cost function $c_i(x^c, \tau, w)$ in (2) for customized and standardized products, respectively. To ease readability in these two subsections, we suppress the notation depicting the dependence of the fluid processes and policies on a fixed total workload w. Then we perform the cost minimization in §3.3 and translate the solution into a dynamic cyclic policy for the original queueing control problem in §3.4.

3.1. Customized Products. The primary challenge in calculating $c_i(x^c, \tau, w)$ is to determine how due-dates affect this cost function in the fluid limit. We begin by normalizing the due-date lead times. Because queue lengths and waiting times are typically $O(\sqrt{n})$ in heavy traffic models, it is appropriate for due-date lead times to also

be $O(\sqrt{n})$. Therefore, we assume that the due-date lead time density $\tilde{f}_i(\sqrt{n}s)$ converges to the nontrivial density $f_i(s)$, with cumulative distribution function $F_i(s) = \tilde{F}_i(\sqrt{n}s)$ and mean $\bar{l}_i = \tilde{l}_i/\sqrt{n}$. In contrast, under a diffusion time scaling the density $\tilde{f}_i(ns)$ converges to a point mass at $s = 0$ and is zero elsewhere. Hence, due-dates do not appear in the diffusion process and are isolated in the fluid limit. This state of affairs is consistent with the heavy traffic snapshot principle (see Reiman), which states that a customer's sojourn in the system is instantaneous under the diffusion scaling.

Let $L_i(t) = \tilde{L}_i(\sqrt{n}t)/\sqrt{n}$ denote the minimum slack process in the fluid system, and define the process $\tilde{D}_i(s,t)$ to be the amount of product i work present in the system at time t that is due at time $s+t$; in other words, $\tilde{D}_i(s,t)$ is the work present at time t that has slack s. Let $\bar{D}_i(s,t) = \tilde{D}_i(\sqrt{n}s, \sqrt{n}t)/\sqrt{n}$ denote its fluid limit. Notice that $D_i(s,t)$ provides a more detailed description of the orders than $W_i(t)$.

Equation (1) shows that the key to calculating the cost function is to determine the behavior of the minimum slack $L_i(t)$ throughout the course of the cycle. The cost function $c_i(x^c, \tau, w)$ is derived in three steps (in Proposition 2, Proposition 3 and equation (11), respectively): We calculate the process $\bar{D}_i(s,t)$ in terms of the original problem parameters and the minimum slack at the start of the cycle, derive $\bar{L}_i(t)$ in terms of the process $\bar{D}_i(s,t)$ (in these first two steps, we replace $\bar{D}_i(s,t)$ by a process closely related to it, as explained after Proposition 1), and express the cost function $c_i(x^c, \tau, w)$ in terms of $\bar{L}_i(t)$.

Without loss of generality, we assume that a cycle for product i starts at time zero and service is received from time $(1-\rho_i)\tau$ to τ. Hence, the amount of product i work in the system at the start of a cycle, which we denote by x_i^s, is equal to $x_i^c - \tau\rho_i(1-\rho_i)/2$. The initial work x_i^s must be stored in the system as orders with due-date lead times greater than or equal to $\bar{L}_i(0)$, by definition the smallest slack among product i orders in the system at time zero. The relation between $\bar{D}_i(s,t)$ and the workload $W_i(t)$ at time zero is given by the following proposition, where $F_i^c(s)$ is defined as $1 - F_i(s)$.

Proposition 1 *At time $t = 0$,*

$$\bar{D}_i(s,0) = \begin{cases} \rho_i \bar{F}_i^c(s) & \text{if } s \geq \bar{L}_i(0) \\ 0 & \text{if } s < \bar{L}_i(0) \end{cases} \tag{4}$$

and

$$x_i^s = \int_{\bar{L}_i(0)}^{\infty} \rho_i \bar{F}_i^c(s) ds. \tag{5}$$

Proof: Because the product i workload arrival process is deterministic and flows in at rate ρ_i under the fluid scaling, at any dt instant of time $\bar{f}_i(s)\rho_i dt$ units of work due in s units of time arrive in the fluid limit. The work in the system at time t with slack s is bounded by the maximum amount of work that could have arrived with a due-date of $t+s$. At time zero, the maximum amount of work present with slack s is the recently arrived work plus work that arrived r units in the past with a due-date of $s+r$. Notationally, this is $\int_0^{\infty} \rho_i \bar{f}_i(s+r)dr$, which is equal to $\rho_i \bar{F}_i^c(s)$. Thus, $\bar{D}_i(s,0) \leq \rho_i \bar{F}_i^c(s)$. This inequality is strict if orders due after time $\bar{L}_i(0)$ were worked on. Since an earliest due-date policy is being used and the machine processes work at rate one, which is strictly greater than $\rho_i \bar{F}_i^c(s)$, it follows that $\bar{D}_i(s,t)$ at the next instant either vanishes or is untouched by service and only affected by arrivals. At time zero, the machine has just switched out of producing product i, implying that the work due after time $\bar{L}_i(0)$ has not been touched and that there has been no opportunity for orders to arrive with due-date lead times below $\bar{L}_i(0^-)$, the smallest slack the instant before the switchover. Thus, equation (4) holds. By construction, we have $x_i^s = \int_{\bar{L}_i(0)}^{\infty} \bar{D}_i(s,0)ds$. Combining this with equation (4) yields equation (5). ∎

Because the interaction of $\bar{D}_i(s,t)$ and $\bar{L}_i(t)$ is complex, we streamline our calculations by creating a simplified variant of $\bar{D}_i(s,t)$ that evolves through the course of one cycle. Let $\bar{D}_i^N(s,t)$ be the amount of work at time t with slack s *if the machine performs no work for product i from $t = 0$ until τ*: $\bar{D}_i^N(s,t)$ is only defined for $t \in [0,\tau)$. Thus, $\bar{D}_i^N(s,t) = \bar{D}_i(s,t)$ for $s \geq \bar{L}_i(t)$ but is not necessarily 0 for $s < \bar{L}_i(t)$. The process $\bar{D}_i^N(s,t)$ is useful for two reasons: Its behavior is easy to describe, and $\bar{L}_i(t)$ can be derived directly from it. The evolution of $\bar{D}_i^N(s,t)$ is described by the following proposition.

Proposition 2 *For $t \in [0, \tau)$,*

$$D_i^N(s,t) = \begin{cases} \rho_i F_i^c(s) & \text{if } s \geq L_i(0) - t \\ \\ \rho_i (\bar{F}_i^c(s) - \bar{F}_i^c(s+t)) & \text{if } s < L_i(0) - t \end{cases} . \tag{6}$$

Proof: By construction, $D_i^N(s,t)$ evolves according to the differential equation

$$\bar{D}_i^N(s, t+dt) = \bar{D}_i^N(s+dt, t) + \rho_i \bar{f}_i(s). \tag{7}$$

That is, the amount of product i work in the system at time $t + dt$ that is due at time $t + s + dt$ is equal to the amount of work that was previously in the system at time t with slack $s + dt$ *plus* the amount of work with a due-date lead time of s that has just arrived. Using the fact that $\bar{D}_i^N(s,t)$ equals $D_i(s,t)$ at $t = 0$, readers can verify that

$$D_i^N(s,t) = \rho_i (F_i^c(s) - \bar{F}_i^c(s+t)) + \bar{D}_i(s+t, 0) \tag{8}$$

is the solution to equation (7). The proposition follows by using equation (4) to substitute in for $\bar{D}_i(s+t, 0)$ in (8). ∎

In an attempt to enhance the reader's intuition, we point out some noteworthy features regarding the evolution of $\bar{D}_i^N(s,t)$ in Figure 2. For the sake of concreteness, Figure 2a contains a uniform due-date lead time distribution, with fluid minimum and maximum of $\bar{a}_i = a_i/\sqrt{n}$ and $\bar{b}_i = b_i/\sqrt{n}$. Figures 2b-2d display, for a fixed value of t, $D_i^N(s,t)$ as a function of s under three cases depending on the relative value of $L_i(0)$, $\bar{a}_i$ and $\bar{a}_i + t$. Although this figure represents snapshots of $\bar{D}_i^N(s,t)$ at a fixed time t, it is perhaps more instructive to describe the dynamics of $D_i^N(s,t)$ as t increases. As time t progresses through the cycle, the existing orders age and new orders arrive. A useful metaphor is to imagine the area under the curves in Figures 2b-2d as water, the curves as waves, and new orders as raindrops accumulating on the waves. Then the waves of work travel to the left (i.e., slack is decreasing) in Figures 2b-2d as time progresses, and grow in height as new orders are added to them. Notice that work with slack $L_i(0)$ at time zero has slack $L_i(0) - t$ at time t if no work is performed during the cycle. By

equation (6), for slacks s exceeding this critical value, the work at time t that has slack s is $\bar{D}_i^N(s,t) = \rho_i \bar{F}_i^c(s)$. For these slack values, which correspond to the regions in Figures 2b-2d to the right of $\bar{L}_i(0) - t$, the quantity $D_i^N(s,t)$ is in *equilibrium*: The amount of work that is aging equals the amount of work that is arriving for all values of $s \geq \bar{L}_i(0) - t$, and so the shape of the waves in Figures 2b-2d to the right of the barrier $L_i(0) - t$ are invariant over time t. For a fixed time t in these equilibrium regions, $D_i^N(s,t)$ is proportional to the complementary cumulative distribution function of the due-date lead time, and so $\bar{D}_i^N(s,t)$ drops to zero at the point $s = \bar{b}_i$. Also, for $s \in [\bar{L}_i(0) - t, a]$, $\bar{D}_i^N(s,t)$ is constant and takes on its maximum value of ρ_i (see Figures 2b and 2c); here, all of the work that could be due at time $s + t$ has arrived.

By Proposition 1, at time $t = 0$, $D_i^N(s,0)$ has the equilibrium value $\rho_i F_i^c(s)$ for $s \geq \bar{L}_i(0)$ and 0 otherwise. However, the evolution over the cycle of $\bar{D}_i^N(s,t)$ has two different qualitative structures depending on if the earliest possible arriving due-date lead time a_i is greater than the initial minimum slack $\bar{L}_i(0)$, or less than it. If $\bar{L}_i(0) < \bar{a}_i$ then no new orders will have a slack less than those that had a slack of $\bar{L}_i(0)$ at time zero. Therefore, orders arrive with just the correct due-date lead time distribution to maintain the equilibrium behavior of $D_i^N(s,0)$, as displayed throughout Figure 2b. However, if $\bar{L}_i(0) > a_i$ then new orders can arrive with a due-date lead time smaller than the minimum slack of the orders present at time zero. In this case, as time t progresses, a small amount of orders with slacks less than $\bar{L}_i(0) - t$ accumulates but never exceeds $\rho_i F_i^c(L_i(0) - t)$, which is the quantity of work in the system at time t with due-date leadtime $\bar{L}_i(0) - t$ (see Figure 2d). Thus, the critical slack value of $\bar{L}_i(0) - t$ marks the *barrier* between the equilibrium work level $\rho_i \bar{F}_i^c(s)$ (to the right of the barrier) and the accumulation of new orders with small due-date lead times (to the left of the barrier). Finally, referring to the portion of the curve to the left of the barrier in Figures 2b-2d, in the region $s < \bar{L}_i(0) - t$, all of the work in $L_i(0)$ is due after time $s + t$, and hence $\bar{D}_i^N(s,t)$ must equal zero for $s < a_i - t$: this critical value of s corresponds to an order arriving at time zero with the minimum due-date lead time of a_i.

Aldous, Kelly and Lehoczky (1995) use heavy traffic theory to analyze the performance of a single-class $GI/G/1$ queue with random due-date lead times that operates

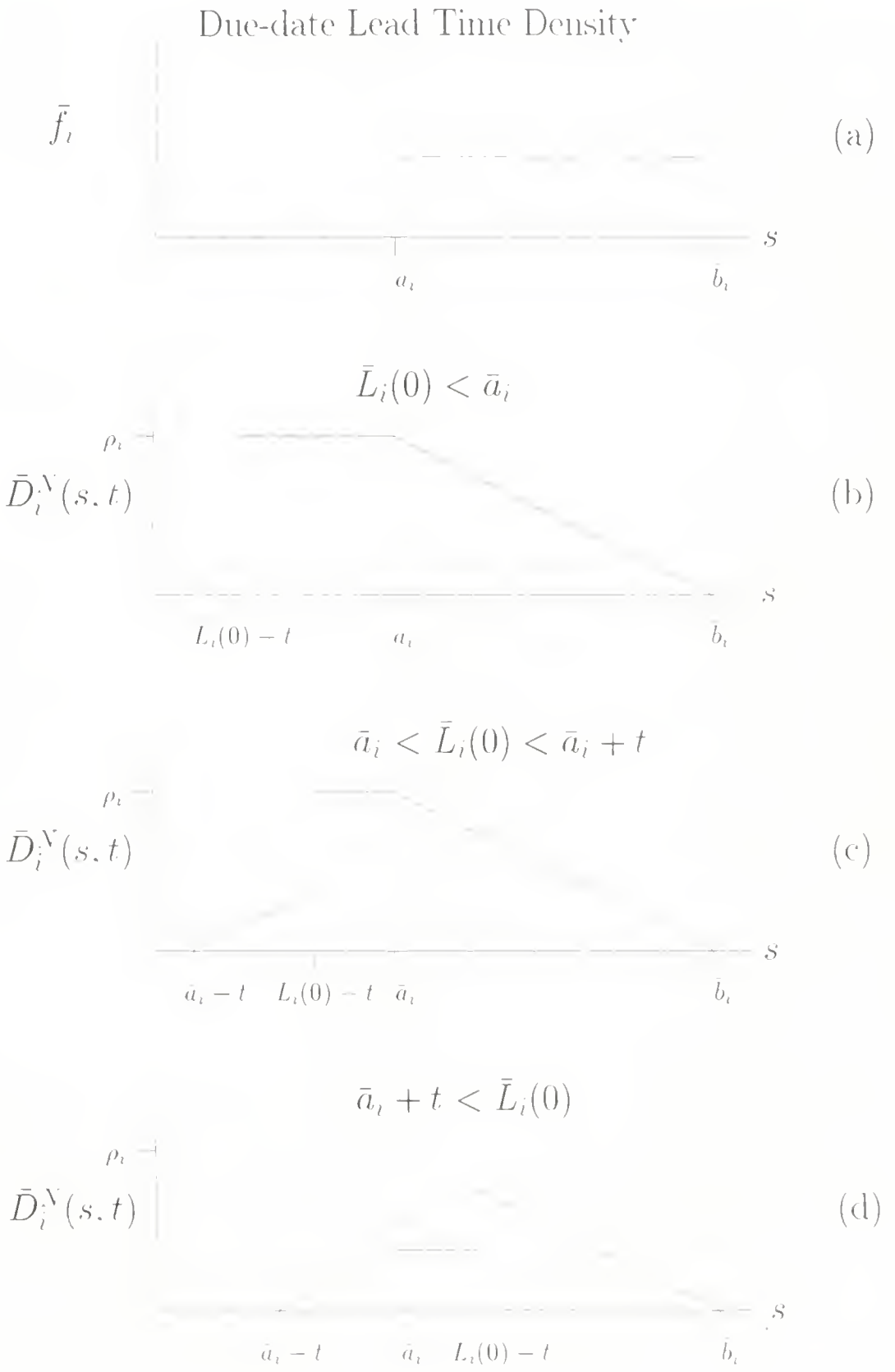

Figure 2: The function $D_i^N(s,t)$ as a function of s.

under the first-come first-served discipline. They derive a figure that is essentially identical to Figure 2b. Because there are no setups and only one product, the non-equilibrium region to the left of the due-date lead time barrier $\bar{L}_i(0) - t$ does not appear in their problem.

Our next step is to calculate the minimum slack process $\bar{L}_i(t)$ in terms of $\bar{D}_i^N(s,t)$. When the machine is servicing other products, the order with the earliest due-date corresponds either to the earliest due-date request just as the server switched out of product i (i.e. $\bar{L}_i(0) \leq \bar{a}_i$) or the request with the due-date lead time of $\bar{a}_i$ that just arrived after the machine switched out (i.e., $\bar{L}_i(0) > a_i$). Thus, we have

$$\bar{L}_i(t) = \min[\bar{a}_i - t, \bar{L}_i(0) - t] \quad \text{for} \quad t \in [0, \tau(1-\rho_i)). \tag{9}$$

Recall that $\bar{D}_i^N(s,t)$ was constructed under the assumption that the machine does not work on product i throughout the cycle. When product i is being served, the machine consumes the left most part of the curves in Figures 2b-2d (in the context of our metaphor, the machine swallows the water at a constant rate). For $t \in (\tau(1-\rho_i), \tau)$, the server would have completed $t - \tau(1-\rho_i)$ units of work. Since the machine works according to the earliest due-date rule within each product, work corresponding to $\bar{D}_i^N(s,t)$ for s below $\bar{L}_i(t)$ has been completed. Thus we have the following proposition.

Proposition 3 *$\bar{L}_i(t)$ is the smallest quantity that satisfies*

$$t - \tau(1-\rho_i) = \int_{-\infty}^{\bar{L}_i(t)} \bar{D}_i^N(s,t)ds \quad \textit{for} \quad t \in (\tau(1-\rho_i), \tau). \tag{10}$$

Equations (5), (9) and (10) uniquely determine $\bar{L}_i(t)$ over the course of one cycle.

Equations (5), (9) and (10) can be used to express the minimum slack process $\bar{L}_i(t)$ over the course of one cycle solely in terms of the initial minimum slack and the primitive probabilistic processes of the problem (although we do not write out this expression here). In addition, we can summarize the qualitative behavior of the minimum slack process. From time zero until time $\tau(1-\rho_i)$ no services occur and only orders arrive to the queue. Since orders age and get closer to their due-date, the minimum slack in product i, $\bar{L}_i(t)$,

is monotonically decreasing at unit rate for $t \in (0, \tau(1-\rho_i))$, as seen in (9). From time $\tau(1-\rho_i)$ until τ, $\bar{L}_i(t)$ is monotonically increasing because the service rate is always greater than $\bar{D}_i(s,t)$ (i.e., orders are being filled faster than they can age), although its behavior is more complex. If $\bar{L}_i(0) < \bar{a}_i$ then, when the machine begins working on product i, $\bar{L}_i(t)$ increases linearly until the end of the cycle; referring to Figure 2b, the machine is consuming the flat part of the curve and $\bar{L}_i(t)$ does not reach a_i before the end of the cycle. If $\bar{L}_i(0) > \bar{a}_i$ then the rate of increase is volatile, moving quickly through the region to the left of the barrier $\bar{L}_i(0) - t$ in Figures 2c and 2d, dramatically slowing when the barrier is passed and then speeding up again as the right tail (represented by $\rho_i F_i^c(s)$ in Figures 2c and 2d) is reached.

Our final step is to determine product i's average inventory cost per cycle in terms of $\bar{L}_i(t)$ for a given cycle center x_i^c and cycle length τ. It is equal to the average of the earliness or tardiness costs associated with orders as they are filled. Since the machine follows an earliest due-date policy, the earliness or tardiness of an order filled at time t is either $\bar{L}_i(t)$ or the due-date lead time associated with a current arrival if that arrival has a due-date lead time less than $\bar{L}_i(t)$. If there are arrivals with due-date lead times s less than $\bar{L}_i(t)$, then the machine spends $\rho_i \bar{f}_i(s)$ fraction of effort on them and $1 - \rho_i \bar{F}_i(\bar{L}_i(t))$ fraction of effort on orders with slack $\bar{L}_i(t)$. Thus the average cost for a product i order is

$$c_i(x_i^c, \tau, w) = \frac{1}{\tau} \int_{\tau(1-\rho_i)}^{\tau} \left(b_i L_i^-(t) + h_i \left[\left(1 - \rho_i \bar{F}_i(\bar{L}_i(t))\right) L_i^+(t) + \int_0^{\bar{L}_i(t)} \rho_i \bar{f}_i(s) s ds \right] \right) dt. \tag{11}$$

Although this cost function is complex, the average cost per cycle is computable. The HTAP has dramatically simplified the problem. As noted in §1, the state of the system in a Markov decision process framework is unwieldy because the evolution of orders with due-dates needs to be tracked over time. The fluid limit has transformed order progression through this complex state space into $\bar{D}_i(s,t)$. From a functional analysis point of view, the ideas are similar because $D_i(\cdot, t)$ is a bounded function on a compact domain and so is a point in the infinite-dimensional space of square integrable functions L^2. The fluid limit thus approximates the evolution of orders in the system as a path

in L^2, parameterized by the index t in $D_i(\cdot, t)$. Although this relationship is abstract, the path in L^2 is made computationally tractable by Propositions 1-3. Moreover, by "reversing the order of integration," we are able to take advantage of this tractability and translate $D_i(s,t)$ into an average cost per cycle.

3.2. Standardized Goods. The cost calculations for standardized products are more complex than in the customized case because the queue of orders and the inventory of completed items are embedded in the products' workload. Let $\bar{W}_i^I(t)$ represent the amount of product i work in finished goods inventory at time t (i.e., the amount of machine time embedded in $\bar{I}_i(t)$), and let $W_i^I(t) = \bar{W}_i^I(\sqrt{n}t)/\sqrt{n}$ be its fluid counterpart. An important aspect of $W_i^I(t)$ is that it must be nonnegative as it represents actual goods in inventory; backorders are in the form of unfilled orders in $D_i(s,t)$. We assume, as in the customized case, that at time zero the server has just switched out of product i. Because $D_i^N(s,t) = D_i(s,t)$ for $s \geq \bar{L}_i(t)$, the workload process $\bar{W}_i(t)$ can be expressed as

$$\bar{W}_i(t) = \int_{\bar{L}_i(t)}^{\infty} D_i^N(s,t)ds - \bar{W}_i^I(t) \quad \text{for} \quad t \in [0,\tau). \tag{12}$$

Recall that we do not assign completed units to standardized orders before their due-date. This leads to the important observation that $L_i(t) \leq 0$ for all t. The rational for this is simple: If for some unexplainable reason $\bar{L}_i(t) > 0$ (for instance when a rare event suddenly shifts the total workload level $\bar{W}_i(t)$) then we assign no completed units to orders and instead place them in finished goods inventory. The minimum slack $L_i(t)$ then decreases to zero and never again goes higher. This is a transitory effect that will be washed away after the repetition of several cycles and so can be ignored. Hence, by equation (6) we can conclude that

$$D_i^N(s,t) = \rho_i F_i^c(s) \quad \text{for} \quad s \geq 0 \quad \text{and} \quad t \in [0,\tau). \tag{13}$$

Because units from inventory are allocated to prevent backorders, it follows that $L_i(t) < 0$ only when $\bar{W}_i^I(t) = 0$; i.e.,

$$W_i^I(t)L_i(t) = 0 \quad \text{for all} \quad t. \tag{14}$$

So if $L_i(t) < 0$ then the next unit of product i completed by the machine is assigned to the product i order with the earliest due-date, and a tardiness cost is incurred. Completed units that are not assigned to orders are placed in the finished goods inventory represented by $\bar{W}_i^I(t)$ and accumulate a holding cost. Therefore, the cost per cycle for a standardized product is

$$c_i(x_i^c, \tau, w) = \frac{1}{\tau}\left[\int_{\tau(1-\rho_i)}^{\tau} b_i \bar{L}_i^-(t)dt + \int_0^{\tau} h_i \bar{W}_i^I(t)dt\right]. \tag{15}$$

Equations (12)-(14) lead to the useful relation that $\bar{W}_i^I(t)$ is positive only if $\bar{W}_i(t) < \int_0^{\infty} \rho_i F_i^c(s)ds$, and so $\bar{W}_i^I(t) = \left(\bar{W}_i(t) - \rho_i \bar{l}_i\right)^-$. It is important to note that we already know the behavior of $\bar{W}_i(t)$ for standardized products from MRW. Thus, the holding cost portion of the cost per cycle in (15) is just a translation in terms of workload (or equivalently, cycle center x_i^c) by $\rho_i \bar{l}_i$ of the stochastic economic lot scheduling problem (SELSP) holding cost.

Now we turn to the tardiness costs in (15). Because $L_i(t) < 0$ when tardiness costs are incurred, it follows by (12)-(14) that

$$\bar{W}_i(t) = \int_{\bar{L}_i(t)}^{0} \bar{D}_i^N(s,t)ds + \int_0^{\infty} \rho_i \bar{F}_i^c(s)ds \tag{16}$$

during these times. To determine $\bar{D}_i^N(s,t)$ in the first integral, we observe that $\bar{L}_i(t) \geq L_i(0) - t$ by (9) and the fact that $\bar{L}_i(t)$ is nondecreasing for $t \in (\tau(1-\rho_i), \tau)$. Hence, $s \geq L_i(0) - t$ in the first integral, and so $\bar{D}_i^N(s,t) = \rho_i F_i^c(s)$ by Proposition 2. Because $s \leq 0$ in this integral, equation (16) becomes $\bar{W}_i(t) = \int_{\bar{L}_i(t)}^{0} \rho_i ds + \rho_i \bar{l}_i$. We conclude that in the backorder regions

$$\rho_i \bar{L}_i^-(t) = \left(\bar{W}_i(t) - \rho_i \bar{l}_i\right)^+. \tag{17}$$

By equations (15) and (17), the time average tardiness cost for a given total workload is

$$\frac{b_i}{\rho_i \tau}\left[\int_{\tau(1-\rho_i)}^{\tau} \left(\bar{W}_i(t) - \rho_i \bar{l}_i\right)^+ dt\right]. \tag{18}$$

Again, this is a translation of the SELSP cost per cycle.

Therefore, the cost per cycle for standardized goods with due-dates is exactly the same as the SELSP cost with a cycle center shift by $\rho_i \bar{l}_i$, which is just product i's

utilization times its mean due-date lead time. For a system with only standardized products, it follows that the optimal switching curves are shifted by $\rho_i \bar{l}_i$ from the SELSP case, and the cost is independent of the due-date lead times; this important observation is discussed further in §4.5, §5.6 and §5.8. Thus, as in MRW, $c_i(x^c, \tau, w)$ is broken down into three regions based on if there is only holding, only backordering or mixed costs over the cycle. The cost per cycle can be expressed as

$$c_i(x^c, \tau, w) = \begin{cases} h_i(\rho_i \bar{l}_i - x_i^c) & \text{if } \rho_i \bar{l}_i - x_i^c > \frac{\tau \rho_i (1-\rho_i)}{2} \\ (b_i + h_i)\frac{\tau \rho_i (1-\rho_i)}{8} + \frac{h_i - b_i}{2}(\rho_i \bar{l}_i - x_i^c) & \text{if } 0 \in [\rho_i \bar{l}_i - x_i^c \pm \frac{\tau \rho_i (1-\rho_i)}{2}] \\ \quad + \frac{b_i + h_i}{2\tau \rho_i (1-\rho_i)}(\rho_i \bar{l}_i - x_i^c)^2 & \\ -b_i(\rho_i \bar{l}_i - x_i^c) & \text{if } \rho_i \bar{l}_i - x_i^c < -\frac{\tau \rho_i (1-\rho_i)}{2} \end{cases} . \tag{19}$$

3.3 Optimization. With an expression for average cost for each level of workload in hand, we can optimize over the dynamic cyclic policy as determined by x^c, τ and w_0. The generality of the due-date lead time distribution prevents an exact solution. A numerical procedure, however, is possible. Policy optimization must be performed on both the fluid and the diffusion levels.

Under the fluid scalings, the cycle center x^c can be optimized with respect to a given cycle length τ and total workload level w. This nonlinear program can be stated as follows:

$$\min_{x^c \in \Re^N} \quad \sum_{i=1}^N c_i(x^c, \tau, w) \tag{20}$$

$$\text{such that:} \quad \sum_{i=1}^N x_i^c = w \tag{21}$$

$$x_i^c \geq \frac{\tau \rho_i (1-\rho_i)}{2} \quad \text{for } i = 1, \ldots, N^c. \tag{22}$$

The highly nonlinear aspects of $\bar{L}_i$ and thus $c_i(x^c, \tau, w)$ make this problem complex. Nonetheless, the problem can be solved numerically using standard descent methods, and we denote the solution by $x^{c^*}(\tau, w)$.

In the diffusion scheme, the long run average cost of the entire problem is minimized. Given that the optimal cycle center $x^{c^*}(\tau, w)$ is known, the minimization of equation (3

reduces to a diffusion control problem with a drift control (via $\tau(w)$) and a singular control (via w_0). Let $V(w)$ be the potential (relative value) function and g be the gain (optimal long run average cost). Although the drift of W is unbounded, we assume that standard arguments apply (see Mandl 1968), and express the associated Hamilton-Jacobi-Bellman optimality equations, after using equation (2), as

$$\min_{\tau(w)} \left\{ \sum_{i=1}^{N} c_i(x^{c^*}(\tau, w), \tau, w) + \frac{K + sV'(w)}{\tau} - g - cV'(w) + \frac{\sigma^2}{2} V''(w) \right\} = 0 \quad \text{for } w \leq w_0 \tag{23}$$

and

$$V'(w) = 0 \quad \text{for } w \geq w_0. \tag{24}$$

As in MRW, we numerically solve the diffusion control problem by approximating the diffusion process by a Markov chain. This Markov chain approximation algorithm was developed by Kushner (1977), and we refer readers to Kushner and Dupuis (1992) for details of the algorithm. MRW contains a full description of its application to solving the corresponding optimality equations for the SELSP. Because the optimality equations in MRW are very similar to equations (23)-(24), we omit a description of the algorithm, and refer readers to Appendix 1.

When there are no setup times, however, the diffusion process W becomes a RBM and the diffusion control problem is easy to solve. Although the cycle length τ and cycle center x^c must still be optimized for each total workload level, they can be done individually without the need to refer to or update the potential function $V'(w)$. Since the steady state distribution of the total workload process W is exponential when setup times are zero, the optimal idling threshold is given by

$$w_0^* = \text{argmin}_{w'} \int_{w'}^{\infty} c(w) \frac{2c}{\sigma^2} e^{-\frac{2c}{\sigma^2}(w-w')} dw, \tag{25}$$

where c and σ^2 are the parameters for the RBM W.

3.4. Proposed Policy. The solution outlined in §3.3 needs to be unscaled to be implemented. Although the translation to an unnormalized policy has considerable room for interpretation, we propose an intuitive policy. The heavy traffic analysis finds

a minimum level of work experienced by an individual product over the course of a cycle for every total workload level. Because work is depleted while the machine is serving it, we propose that the machine continue production on the current product until this minimum amount of work is reached, and then begin setting up for the next product in the cycle. In addition, the machine idles when the total workload level has fallen to the idling threshold.

We define the policy in terms of two quantities that are naturally observed in practice: $\tilde{O}_i(t)$, which is the number of orders of product i in queue for $i = 1, \ldots, N$ and $\tilde{I}_i(t)$, which is the number of completed units of product i in finished goods inventory for $i = N^c + 1, \ldots, N$. For ease of exposition, we let $\tilde{I}_i(t) = 0$ for $i = 1, \ldots, N^c$. The linear identity $O_i - I_i = \mu_i W_i$, where $O_i(t) = \tilde{O}_i(nt)/\sqrt{n}$ and $I_i(t) = \tilde{I}_i(nt)/\sqrt{n}$, which is known to hold for a wide variety of queueing systems in heavy traffic, is used to translate the heavy traffic solution to a scheduling policy in terms of $\tilde{O}_i(t)$ and $\tilde{I}_i(t)$. The proposed policy is: Switch out of product i when

$$\tilde{O}_i(t) - \tilde{I}_i(t) \leq \sqrt{n}\mu_i \left[x_i^{c*}\left(\frac{\sum_{j=1}^{N} \mu_j^{-1}(\tilde{O}_j(t) - \tilde{I}_j(t))}{\sqrt{n}}\right) - \frac{\tau^*\left(\frac{\sum_{j=1}^{N} \mu_j^{-1}(\tilde{O}_j(t) - \tilde{I}_j(t))}{\sqrt{n}}\right)\rho_i(1-\rho_i)}{2} \right] \tag{26}$$

and idle the machine when

$$\sum_{i=1}^{N} \mu_i^{-1}(\tilde{O}_i(t) - \tilde{I}_i(t)) = \sqrt{n} w_0^*, \tag{27}$$

where $x_i^{c*}(w)$, $\tau^*(w)$ and w_0^* are determined from the optimization procedure outlined in §3.3. Because a computational procedure is being employed, we must choose a value of the heavy traffic parameter n, and we let $n = (1-\rho)^{-2}$. As in MRW, exploratory computations (not shown here) reveal that the performance of the proposed policy is very insensitive to our choice of n.

Although we do not do so here, more complicated policies can be created. At any point in time over the cycle, our analysis yields the fluid workload level of each product. This type of information can be used to determine when random events throw the cycle

off course in a manner not predicted by the HTAP's $O(\sqrt{n})$ fluid scaling. The dynamic cyclic approach can thus provide "red flags" or warnings for unusual surges in demand or slow-downs in machine processing. A more complicated policy using this information would specify when to skip products in the cycle or when to jump to a product that has an unusually high number of orders. We leave the specification of such policies for future investigation.

4. DETERMINISTIC DUE-DATE LEAD TIMES

In this section we assume that due-date lead times for each product are deterministic, thereby allowing us to write a closed form expression for the cost per cycle in equation (2) in §4.1. In addition, we are able to state a quick method in §4.2 for determining the optimal cycle center x^{c*} and provide a formulaic solution. With this, cycle length τ is found in terms of $V'(w)$ and the total workload w. However, the solution to the diffusion control problem must still be computed by the Markov chain approximation algorithm. The proposed solution is stated in §4.3. Subsections 4.4 and 4.5 are devoted to some structural properties of the optimal solution and the value of due-date lead times, respectively.

4.1 Cost per Cycle. Let each product i have a deterministic due-date lead time of $\tilde{f}_i$, which is $f_i = \tilde{f}_i/\sqrt{n}$ under the fluid scaling. The cumulative distribution function is

$$F_i^c(s) = \begin{cases} 0 & \text{if } s > f_i \\ 1 & \text{if } s \le f_i \end{cases} . \tag{28}$$

For a given cycle center x_i^c, cycle length τ and total workload level w, the cost per cycle for a standardized product is given by (19), with f_i in place of l_i.

To find the cost per cycle for a customized product, we must determine the behavior of $\bar{D}_i(s,t)$, $\bar{L}_i(t)$ and $\bar{D}_i^N(s,t)$ given x_i^c and τ. By Proposition 1, we have $x_i^s = \int_{\bar{L}_i(0)}^{f_i} \rho_i ds$, or $\bar{L}_i(0) = f_i - x_i^s/\rho_i$. Therefore, $\bar{L}_i(0)$ has a range of $[-\infty, f_i]$, since for customized products x_i^s is greater than or equal to zero. This makes intuitive sense: The minimum slack must be no larger than f_i because no orders can arrive with larger due-date lead

times. From Propositions 1 and 2 respectively, we have

$$D_i(s,0) = \begin{cases} \rho_i & \text{if } s \in (\bar{f}_i - x_i^s/\rho_i, f_i) \\ 0 & \text{otherwise} \end{cases} \tag{29}$$

and

$$D_i^N(s,t) = \begin{cases} 0 & \text{if } s \le \bar{f}_i - x_i^s/\rho_i - t \\ \rho_i & \text{if } \bar{f}_i - x_i^s/\rho_i - t \le s \le f_i \\ 0 & \text{if } \bar{f}_i < s \end{cases} . \tag{30}$$

We can solve for $L_i(t)$ by using Proposition 3: For $t \in (\tau(1-\rho_i), \tau)$,

$$\bar{L}_i(t) = \bar{f}_i - x_i^s/\rho_i - (1-\rho_i)\tau + (1-\rho_i)(t - \tau(1-\rho_i))/\rho_i. \tag{31}$$

Because $(1-\rho_i)(t-\tau(1-\rho_i))/\rho_i - (1-\rho_i)\tau$ is less than or equal to zero, $\bar{L}_i(t)$ is also bounded above by $\bar{f}_i$. Note that this function depends on the the decision variable x_i^c via the initial inventory x_i^s.

We can now substitute (31) into equation (11) to calculate the average cost per cycle. Since no orders arrive with due-date lead times $\bar{L}_i(t)$ less than $\bar{f}_i$, equation (11) can be rewritten as

$$c_i(x^c, \tau, w) = \frac{1}{\tau} \int_{\tau(1-\rho_i)}^{\tau} \left(b_i \bar{L}_i^-(t) + h_i \bar{L}_i^+(t) \right) dt. \tag{32}$$

This structure is similar to that in the SELSP case discussed in MRW. The cost per cycle is broken down into three cases depending on if orders are only tardy, only early or both during the course of the cycle. If $\bar{L}_i(t)$ is always greater than zero over the course of the cycle (i.e., if $\rho_i f_i - x_i^c > \tau\rho_i(1-\rho_i)/2$) then orders are always filled early and so equation (32) implies

$$c_i(x^c, \tau, w) = \frac{1}{\tau} h_i \int_{(1-\rho_i)\tau}^{\tau} \left[\bar{f}_i - x_i^s/\rho_i - (1-\rho_i)\tau + (1-\rho_i)(t - \tau(1-\rho_i))/\rho_i \right] dt, \tag{33}$$

which simplifies to $c_i(x_i^c, \tau, w) = h_i\rho_i(\bar{f}_i - x_i^c/\rho_i)$. This is very similar to the results from MRW: The cost per cycle is equal to that in the SELSP with axes shifted by $\rho_i \bar{f}_i$. The similarity can be viewed as an application of Little's law. In our notation, the utilization ρ_i corresponds to the arrival rate of work and $\bar{f}_i - x_i^c/\rho_i$ is the average waiting time, and

their product is the number in queue.

Similarly, if orders are always tardy then $L_i(t)$ is less than 0 for all $t \in [0, \tau]$, and the cost is $c_i(x^c, \tau, w) = b_i\rho_i[(x_i^c/\rho_i) - \bar{f}_i]$. If orders are both early and tardy over the cycle then $L_i(t)$ is both positive and negative over the cycle and the average cost per cycle is

$$\begin{aligned} c_i(x^c, \tau, w) &= b_i\rho_i \left[\frac{(f_i - \frac{x_i^c}{\rho_i})^2}{2\tau(1-\rho_i)} - \frac{1}{2}(\bar{f}_i - \frac{x_i^c}{\rho_i}) + \frac{1}{8}\tau(1-\rho_i) \right] \\ &\quad + h_i\rho_i \left[\frac{(f_i - \frac{x_i^c}{\rho_i})^2}{2\tau(1-\rho_i)} + \frac{1}{2}(\bar{f}_i - \frac{x_i^c}{\rho_i}) + \frac{1}{8}\tau(1-\rho_i) \right]. \end{aligned} \tag{34}$$

Again, this reduces to a Little's law version of the SELSP results.

For ease of reference, if the parameters are such that $\rho_i\bar{f}_i - x_i^c > \tau\rho_i(1-\rho_i)/2$, we say that product i is in *condition 1*. Similarly, we call product i in *condition 2* if $0 \in [\rho_i\bar{f}_i - x_i^c \pm \tau\rho_i(1-\rho_i)/2]$ and in *condition 3* if $\rho_i\bar{f}_i - x_i^c < -\tau\rho_i(1-\rho_i)/2$. In summary, the average cost per cycle for customized products is

$$c_i(x^c, \tau, w) = \begin{cases} h_i(\rho_i\bar{f}_i - x_i^c) & \text{for } \textit{condition 1} \\ (b_i + h_i)\frac{\tau\rho_i(1-\rho_i)}{8} + \frac{h_i - b_i}{2}(\rho_i\bar{f}_i - x_i^c) & \\ \quad + \frac{b_i + h_i}{2\tau\rho_i(1-\rho_i)}(\rho_i\bar{f}_i - x_i^c)^2 & \text{for } \textit{condition 2} \\ -b_i(\rho_i f_i - x_i^c) & \text{for } \textit{condition 3} \end{cases}. \tag{35}$$

This leads to a remarkable result: Equations (19) and (35) imply that for the deterministic due-date case the cost structure for customized products and standardized ones are *exactly the same*. The only difference between the two types is the restriction that for customized products the cycle center x_i^c cannot be less than $\rho_i(1-\rho_i)\tau/2$. Due-dates have transformed customized products into quasi-standardized ones. We discuss the interpretation of this result in §5.7 and §5.9.

4.2. Optimization. The explicit derivation of average cost per cycle in §4.1 allows us to make further progress in optimizing this system. First, we derive the optimal cycle center and cycle length, and then construct an algorithm for use in the diffusion control problem.

Cycle Center. By our analysis in §4.1, the only difference between the optimization

of the cycle center in (20)-(22) and in MRW is the inclusion of the equation (22) constraints representing the non-negativity restrictions on customized products. In MRW, the solution to the program without inequalities (22) was found exactly. For later reference, we call (20)-(21) the "unconstrained" version of (20)-(22) and denote its solution by x^{c^*u}. As discussed in MRW, the form of the cycle center x^{c^*u} is broken down into three regions depending on whether the objective function is linear or quadratic in the cheapest product at the point x^{c^*u}. We wish to find a similar expression for x^{c^*} because it is used in determining the optimal cycle length τ.

The major complication in solving (20)-(22) is the boundary conditions in equation (22). We exploit the structure of the objective function to determine the structure of $x_i^{c^*}$ and to find which $x_i^{c^*}$'s are binding with respect to the inequality constraints. The objective function is piecewise quadratic with linear edges. It is convex and so if a local minimum exists it will be a global minimum. As per Bertsekas (1995), the Lagrangian associated with (20)-(22) with fixed cycle length τ and total workload level w is

$$L(x^c, \lambda, \mu) = c(x^c, \tau, w) + \lambda \left(\sum_{i=1}^{N} x_i^c - w \right) + \sum_{j=1}^{N^c} \mu_j \left(\frac{\tau \rho_j (1 - \rho_j)}{2} - x_j^c \right) . \tag{36}$$

The Karesh-Kuhn-Tucker necessary conditions state that for local minimum x^{c^*} there exist Lagrangian multipliers λ^* and μ_j^* for $j = 1, \ldots, N^c$ such that

$$\nabla_{x^c} L(x^{c^*}, \lambda^*, \mu^*) = 0, \tag{37}$$

$$\mu_j^* \geq 0, \tag{38}$$

$$\mu_j^* = 0 \quad \forall j \in \Theta^*, \tag{39}$$

where Θ^* is the set of non-binding cycle centers; i.e., products j such that $x_j^{c^*} > \tau \rho_j (1 - \rho_j)/2$. We suppress the dependence of Θ^* on τ and w for increased readability. For additional ease of reference, we categorize the binding products by their *condition*. We let Θ^{*1} be the set of products with binding cycle centers and in *condition 1* and Θ^{*2} be the set of binding products in *condition 2* (no binding product can be in *condition 3*). The products with binding cycle centers are pushed to their limit, in the sense

that no more work can be performed on these products. Equation (37) implies that $\frac{\partial}{\partial x_i^c} c_i(x^{c^*}, \tau, w) + \lambda - \mu_i = 0$ for $i = 1, \ldots, N^c$ and $\frac{\partial}{\partial x_i^c} c_i(x^{c^*}, \tau, w) + \lambda = 0$ for $i = N^c + 1, \ldots, N$. Thus, for $i \in \Theta^*$ and $i = 1, \ldots, N^c$, it follows that $\frac{\partial}{\partial x_i^c} c_i(x^{c^*}, \tau, w) + \lambda = 0$. Therefore, the Karesh-Kuhn-Tucker conditions are the same for the non-binding cycle centers as for those in problem (20)-(22). This fact implies that the cycle center optimality arguments in MRW hold for the non-binding Θ^* products. There it was shown that all but the cheapest earliness (minimum h_i) or tardiness (minimum b_i) product are in *condition 2*. The *condition* of the cheapest product is used to categorize the total feasible workload into 3 regions: Region I, where the cheapest product is in *condition 1*; region II, where the cheapest product is in *condition 2*; and region III, where the cheapest product is in *condition 3*. Here we use the same region terminology to denote when the cycle center and cycle length are such that the cheapest product is in *condition 1*, *2* or *3*. Let θ_b^* be the cheapest tardiness cost product in Θ^* and let θ_h^* be the cheapest earliness cost product. For ease of notation, let θ^* equal θ_b^* if $w > \sum_{i=1}^{N} \bar{f}_i$ and θ_h^* otherwise; the index θ^* corresponds to the "N^{th} product" referred to in MRW. If we let $\hat{w} = w - \sum_{j \notin \Theta^*} \tau \rho_j (1 - \rho_j)/2$ then the non-binding product $i \in \{\Theta^* \setminus \theta^*\}$ cycle center is

$$x_i^{c^*} = \begin{cases} \rho_i \bar{f}_i - \frac{\tau \rho_i (1-\rho_i)}{b_i + h_i} \left[\frac{b_i - h_i}{2} + h_{\theta^*} \right] \\ \qquad \text{if } (\sum_{i \in \Theta^*} \rho_i \bar{f}_i - \hat{w}) - \sum_{i \in \{\Theta^* \setminus \theta^*\}} \bar{x}_i^c > \frac{\tau \rho_{\theta^*} (1-\rho_{\theta^*})}{2} \\ \rho_i f_i - \tau \alpha_i \cdot \gamma_1 - (\sum_{i \in \Theta^*} \rho_i \bar{f}_i - \hat{w}) \alpha_i \cdot \gamma_2 \\ \qquad \text{if } |(\sum_{i \in \Theta^*} \rho_i \bar{f}_i - \hat{w}) - \sum_{i \in \{\Theta^* \setminus \theta^*\}} \bar{x}_i^c| \leq \frac{\tau \rho_{\theta^*} (1-\rho_{\theta^*})}{2} \\ \rho_i \bar{f}_i - \frac{\tau \rho_i (1-\rho_i)}{b_i + h_i} \left[\frac{b_i - h_i}{2} - b_{\theta^*} \right] \\ \qquad \text{if } (\sum_{i \in \Theta^*} \rho_i f_i - \hat{w}) - \sum_{i \in \{\Theta^* \setminus \theta^*\}} \bar{x}_i^c < -\frac{\tau \rho_{\theta^*} (1-\rho_{\theta^*})}{2} \end{cases} , \tag{40}$$

and the cheapest product cycle center is $x_{\theta^*}^{c^*} = \hat{w} - \sum_{i \in \Theta^*} x_i^{c^*}$, where, for $i \in \{\Theta^* \setminus \theta^*\}$, $\gamma_1 = (\ldots, (b_i - h_i)/2 - (b_{\theta^*} - h_{\theta^*})/2, \ldots)^T$, $\gamma_2 = (\ldots, (b_{\theta^*} + h_{\theta^*})/(\rho_{\theta^*}(1 - \rho_{\theta^*})), \ldots)^T$, the vector $\alpha_i = (\ldots, \alpha_{ij}, \ldots)^T$ is defined by

$$\alpha_{ij} = -\frac{\frac{\rho_i(1-\rho_i)}{b_i+h_i} \frac{\rho_j(1-\rho_j)}{b_j+h_j}}{\sum_{l \in \Theta^*} \frac{\rho_l(1-\rho_l)}{b_l+h_l}} \quad \text{for} \quad i \neq j, \quad \text{and} \tag{41}$$

$$\alpha_{ii} = \frac{\rho_i(1-\rho_i)}{b_i+h_i} \frac{\sum_{l\in\{\Theta^*\setminus i\}} \frac{\rho_l(1-\rho_l)}{b_l+h_l}}{\sum_{l\in\Theta^*} \frac{\rho_l(1-\rho_l)}{b_l+h_l}}, \tag{42}$$

and finally where $\bar{x}_i^c = \rho_i f_i - \tau\alpha_i\cdot\gamma_1 - (\sum_{j\in\Theta^*}\rho_j f_j - \hat{w})\alpha_i\cdot\gamma_2$. For $i\in\{\Theta^{*1}\cup\Theta^{*2}\}$, we have $x_i^{c^*} = \tau\rho_i(1-\rho_i)/2$.

It is important to note that when a product i is binding yet its Lagrangian multiplier μ_i is zero, the inclusion or absence of i in Θ^* does not affect the cycle centers of the other products. This can be seen by setting the borderline product's cycle center as derived from equation (40) equal to $\tau\rho_i(1-\rho_i)/2$ and by subsequent algebraic manipulations. In addition, on the border between regions II and I or III the cycle center does not change if using any of the two corresponding expressions in equation (40). We can therefore conclude that the cycle center is a continuous function of the cycle length τ.

With these expressions for the cycle center, we can restate the average cost per cycle in terms of the cycle length τ and the total workload level w. For $i\in\Theta^*\setminus\theta^*$, it is

$$c_i(\tau,w) = \begin{cases} (b_i+h_i)\frac{\tau\rho_i(1-\rho_i)}{8} + \frac{\tau\rho_i(1-\rho_i)}{2(b_i+h_i)}\left[\frac{b_i-h_i}{2}+h_{\theta^*}\right]\left[\frac{h_i-h_i}{2}+h_{\theta^*}\right] & \text{I} \\ (b_i+h_i)\frac{\tau\rho_i(1-\rho_i)}{8} + \frac{\tau(b_i+h_i)}{2\rho_i(1-\rho_i)}(\alpha_i\gamma_1)^2 & \\ \quad + \frac{h_i+h_i}{\rho_i(1-\rho_i)}(\alpha_i\gamma_1)(\alpha_i\gamma_2)(w_{\Theta^*}+\tau\sum_{j\notin\Theta^*}\frac{\rho_j(1-\rho_j)}{2}) & \text{II} \\ \quad + \frac{b_i+h_i}{2\tau\rho_i(1-\rho_i)}(w_{\Theta^*}+\tau\sum_{j\notin\Theta^*}\frac{\rho_j(1-\rho_j)}{2})^2(\alpha_i\gamma_2)^2 & \\ \quad + \frac{h_i-b_i}{2}(\tau\alpha_i\gamma_1+(w_{\Theta^*}+\tau\sum_{j\notin\Theta^*}\frac{\rho_j(1-\rho_j)}{2})\alpha_i\gamma_2) & \\ (b_i+h_i)\frac{\tau\rho_i(1-\rho_i)}{8} + \frac{\tau\rho_i(1-\rho_i)}{2(b_i+h_i)}\left[\frac{b_i-h_i}{2}-b_{\theta^*}\right]\left[\frac{h_i-b_i}{2}-b_{\theta^*}\right] & \text{III} \end{cases}. \tag{43}$$

If we define $\alpha_{\theta^*}\cdot\gamma_1 = -\sum_{i\in\{\Theta^*\setminus\theta^*\}}\alpha_i\gamma_1$ and $\alpha_{\theta^*}\cdot\gamma_2 = 1-\sum_{i\in\{\Theta^*\setminus\theta^*\}}\alpha_i\gamma_2$, then the cost per

cycle for the cheapest product is

$$c_{\theta^*}(\tau, w) = \begin{cases} h_{\theta^*}(w_{\Theta^*} + \tau \sum_{J \notin \Theta^*} \frac{\rho_J(1-\rho_J)}{2}) & \text{I} \\ \quad -h_{\theta^*}\tau \sum_{i \in \Theta^* \setminus \theta^*} \frac{\rho_i(1-\rho_i)}{b_i+h_i}\left[\frac{b_i-h_i}{2} + h_{\theta^*}\right] & \\ (b_{\theta^*} + h_{\theta^*})\frac{\tau\rho_{\theta^*}(1-\rho_{\theta^*})}{8} + \frac{\tau(b_{\theta^*}+h_{\theta^*})}{2\rho_{\theta^*}(1-\rho_{\theta^*})}(\alpha_{\theta^*}\gamma_1)^2 & \\ \quad + \frac{h_{\theta^*}+h_{\theta^*}}{\rho_{\theta^*}(1-\rho_{\theta^*})}(\alpha_{\theta^*}\gamma_1)(\alpha_{\theta^*}\gamma_2)(w_{\Theta^*} + \tau \sum_{J \notin \Theta^*} \frac{\rho_J(1-\rho_J)}{2}) & \text{II} \\ \quad + \frac{b_{\theta^*}+h_{\theta^*}}{2\tau\rho_{\theta^*}(1-\rho_{\theta^*})}(w_{\Theta^*} + \tau \sum_{J \notin \Theta^*} \frac{\rho_J(1-\rho_J)}{2})^2(\alpha_{\theta^*}\gamma_2)^2 & \\ \quad + \frac{h_{\theta^*} - b_{\theta^*}}{2}(\tau\alpha_{\theta^*}\gamma_1 + (w_{\Theta^*} + \tau \sum_{J \notin \Theta^*} \frac{\rho_J(1-\rho_J)}{2})\alpha_{\theta^*}\gamma_2) & \\ -b_{\theta^*}(w_{\Theta^*} + \tau \sum_{J \notin \Theta^*} \frac{\rho_J(1-\rho_J)}{2}) & \text{III} \\ \quad + b_{\theta^*}\tau \sum_{i \in \Theta^* \setminus \theta^*} \frac{\rho_i(1-\rho_i)}{b_i+h_i}\left[\frac{b_i-h_i}{2} - b_{\theta^*}\right]) & \end{cases} . \quad (44)$$

The products with binding cycle centers can be in *condition 1* or in *condition 2*. Thus for $i \in \{\Theta^{*1} \cup \Theta^{*2}\}$ the cost per cycle is

$$c_i(\tau, w) = \begin{cases} h_i(\rho_i \bar{f}_i - \frac{\tau\rho_i(1-\rho_i)}{2}) & \text{for } i \in \Theta^{*1} \\ \frac{b_i+h_i}{2\tau\rho_i(1-\rho_i)}(\rho_i \bar{f}_i)^2 - b_i(\rho_i \bar{f}_i - \frac{\tau\rho_i(1-\rho_i)}{2}) & \text{for } i \in \Theta^{*2} \end{cases} . \quad (45)$$

Cycle Length. Given the form of the cost per cycle in terms of τ and w, we can find an expression for the optimal cycle length τ by differentiating equation (23) with respect to τ and solving against zero. The optimal cycle length is expressed in terms of basic system parameters, the set of binding products and the effective setup cost per cycle $S = K + sV''(w)$. Thus, Θ^*, Θ^{*1} and Θ^{*2} are all functions of S and w. If we define the constants

$$\xi_1^{\Theta^*} = \sum_{i \in \Theta^* \setminus \theta^*} \frac{(b_i+h_i)\rho_i(1-\rho_i)}{8} - \frac{\rho_i(1-\rho_i)}{2(b_i+h_i)}\left(\frac{b_i-h_i}{2} + h_{\theta^*}\right)^2 + \sum_{i \in \Theta^{*1}} \frac{\rho_i(1-\rho_i)}{2}(h_{\theta^*} - h_i) + \sum_{i \in \Theta^{*2}} \frac{\rho_i(1-\rho_i)}{2}(h_{\theta^*} + b_i), \quad (46)$$

$$\xi_2^{\Theta^*} = \sum_{i \in \Theta^{*2}} \frac{(\rho_i \bar{f}_i)^2(b_i+h_i)}{2\rho_i(1-\rho_i)}, \quad (47)$$

$$\xi_3^{\Theta^*} = \sum_{i \in \Theta^*} \left[\frac{(b_i+h_i)\rho_i(1-\rho_i)}{8} + \frac{b_i+h_i}{2\rho_i(1-\rho_i)}(\alpha_i\gamma_1)^2 + \frac{h_i-b_i}{2}\alpha_i\gamma_1 \right. \quad (48)$$

$$+\frac{b_i+h_i}{2\rho_i(1-\rho_i)}(\alpha_i\gamma_2)^2\Big(\sum_{j\notin\Theta^\bullet}\frac{\rho_j(1-\rho_j)}{2}\Big)^2+\frac{h_i-b_i}{2}\alpha_i\gamma_2\Big(\sum_{j\notin\Theta^\bullet}\frac{\rho_j(1-\rho_j)}{2}\Big)\Bigg]$$
$$-\sum_{i\in\Theta^{\bullet 1}}h_i\frac{\rho_i(1-\rho_i)}{2}+\sum_{i\in\Theta^{\bullet 2}}b_i\frac{\rho_i(1-\rho_i)}{2},$$

$$\xi_4^{\Theta^\bullet} = \sum_{i\in\Theta^\bullet}\frac{b_i+h_i}{2\rho_i(1-\rho_i)}(\alpha_i\gamma_2)^2, \quad \text{and} \tag{49}$$

$$\xi_5^{\Theta^\bullet} = \sum_{i\in\Theta^\bullet\setminus\theta^\bullet}\frac{(b_i+h_i)\rho_i(1-\rho_i)}{8}-\frac{\rho_i(1-\rho_i)}{2(b_i+h_i)}\left(\frac{b_i-h_i}{2}-b_{\theta^\bullet}\right)^2 \tag{50}$$
$$-\sum_{i\in\Theta^{\bullet 1}}\frac{\rho_i(1-\rho_i)}{2}(b_{\theta^\bullet}-h_i)+\sum_{i\in\Theta^{\bullet 2}}\frac{\rho_i(1-\rho_i)}{2}(b_i-b_{\theta^\bullet}),$$

then the optimal cycle length can be stated as

$$\tau^* = \begin{cases} \sqrt{\frac{S+\xi_2^{\Theta^\bullet}}{\xi_1^{\Theta^\bullet}}} & \text{I} \\ \sqrt{\frac{S+\xi_4^{\Theta^\bullet}\left(\sum_{i\in\Theta^\bullet}\rho_i\hat{f}_i-w\right)^2+\xi_2^{\Theta^\bullet}}{\xi_3^{\Theta^\bullet}}} & \text{II} \\ \sqrt{\frac{S+\xi_2^{\Theta^\bullet}}{\xi_5^{\Theta^\bullet}}} & \text{III} \end{cases}. \tag{51}$$

In order to use the Markov chain approximation algorithm discussed in §3.3, we need to be able to find τ^* and x^{Θ^*} for a given w and for varying $V'(w)$. Thus, it is necessary to find the set Θ^* of non-binding products as a function of the setup penalty $S = K+sV'(w)$ and workload w. In the Appendix 2, we construct an algorithm that generates the set of binding customized products as a function of effective setup cost for each total workload level. This algorithm allows the cycle length, cycle center and cycle cost to be calculated. This cost can then be fed into the Markov chain approximation algorithm so that $V'(w)$, and hence the proposed policy, can be computed.

4.3. Proposed Policy. The mapping from the solution of the diffusion control problem to the proposed policy is the same as in §3.4, which implements a switching rule for the machine based on $\tilde{W}_i$, the current workload present in individual orders, via $\tilde{O}_i(t)-\tilde{I}_i(t)\approx\mu_i\tilde{W}_i(t)$. In the presence of deterministic due-dates, however, the analysis in §4.1 allows us to create an alternative switching rule based on the minimum slack of each product, via $\tilde{W}_i(t)\approx\rho_i\left(\hat{f}_i-\tilde{L}_i(t)\right)$. This alternative approach might be easier

to implement in cases where due-date data are more readily available than the levels of order queues and inventories. Markowitz (1996) uses Monte Carlo simulation to compare the two alternatives, and finds that the approach given in §3.4 clearly outperforms the minimum slack method, although the discrepancy between the two policies disappears as the length of the due-date lead time grows; see Markowitz for details.

4.4. Structural Properties. Unfortunately, for $s > 0$, no closed form solution to $V'(w)$ appears possible, even if Θ^* is known for all total workload levels. However, it is possible to derive several structural properties about the derivative of the potential function $V'(w)$, which yields some qualitative insight about the proposed policy. We assume that as $w \to \infty$ the set of binding variables Θ^* stabilizes; i.e., there exists a w' such that for all $w_1, w_2 > w'$, we have $\Theta^*(w_1) = \Theta^*(w_2)$. This assumption follows naturally from the behavior evident in observations 1-8 in the Appendix 2, which demonstrates how the products smoothly become binding with respect to S and w, and the fact that there are only a finite number of possibilities for Θ^*. These structural properties are similar to the ones derived in MRW. Please see Markowitz for details of their proofs.

Property 1. *Let the average setup time per cycle s be greater than zero. If region II conditions hold as $w \to \infty$, then*

$$V'(w) = \frac{2\sqrt{\zeta_3^{\Theta^*}\zeta_4^{\Theta^*}} + \zeta_6^{\Theta^*}}{c} w + o(w), \tag{52}$$

where Θ^* is the set containing the standardized products and the cheapest backorder product (this product could be customized or standardized) and

$$\zeta_6^{\Theta^*} = \frac{\sum_{j \notin \Theta^*} \rho_j(1-\rho_j) + \sum_{j \in \Theta^*} (h_j - b_j)\frac{\rho_j(1-\rho_j)}{b_j+h_j}}{2\sum_{j \in \Theta^*} \frac{\rho_j(1-\rho_j)}{b_j+h_j}}. \tag{53}$$

If region III conditions hold as $w \to \infty$, then

$$V'(w) = -\frac{b_{\theta^*}}{c} w + o(w). \tag{54}$$

Property 2. *If the setup cost per cycle $K = 0$ and all of the products are standardized, then the policy at the idling threshold w_0 satisfies the region II conditions if and only if $h_i = h_j$ for all i and j. If this condition holds then*

$$\tau^*(w_0) = \sqrt{\xi_4/\xi_3}\Big(\sum_{i\in\Theta^*} \rho_i f_i - w_0\Big). \tag{55}$$

If $h_i \neq h_j$ for some i and j then the idling threshold satisfies the region I conditions and $\tau^*(w_0) = 0$.

Property 3. *If average setup time s is greater than zero, region II conditions hold in the limit as $w \to \infty$ if and only if the tardiness costs of all the standardized products are equal and the cheapest tardiness cost among the customized products is equal to or greater than the standardized tardiness cost (or if the tardiness cost of all of the customized products are equal and there are no standardized products).*

Like their counterparts in MRW, these three properties provide insights into the nature of the optimal cycle length. By Proposition 3, if the tardiness costs of the cheapest customized product and of all the standardized products are equal, then from equation (51) and Property 1 the optimal cycle length τ^* grows linearly for large total workload w. If the tardiness costs are not equal then from Property 1 and equation (51), τ^* grows as the square root of total workload. Property 2 describes the optimal cycle length near the idling threshold w_0.

Beyond MRW, these three properties make statements about Θ^*, the set of non-binding product classes. In the limiting case of $w \to \infty$, the optimal cycle length is growing and eventually causes all but potentially one customized product to become binding. By Property 2, around the idling threshold, if $K = 0$ and region I conditions hold, then no products are binding. If region II conditions hold or $K > 0$, then Θ^* is *a priori* difficult to describe.

4.5. The Value of Due-date Lead Times. Based on the analysis of the optimal policy, we can comment on how due-date lead times impact the long run average cost. For systems with standardized goods, due-dates do not influence the optimal costs. By a quick inspection of the cost per cycle in equation (19), we see that modifications in

due-date lead times are offset by a translation of the cycle center. Thus, the optimal cost is not altered by changes in the due-date lead times; we describe the intuition behind this result in §5.6.

The situation for customized products is more complex. The translation of cycle centers influences the costs associated with binding products. As due-date lead times lengthen, the feasible regions for the cycle center x^c and cycle length τ grow, thereby allowing for a lower long run average cost. However, for the zero setup time case, we can show that when due-date lead times become sufficiently large, no products are binding and hence the long run average cost is independent of the due-date lead times. We do this by solving the control problem for the $s = 0$ case, both ignoring binding constraints (i.e., treating each customized good as a standardized one so that $\Theta^* = \{1, \ldots, N\}$) and setting the due-date lead times $\bar{f}_i$ equal to zero. Let w_0^0 be the optimal idling threshold. If region II conditions hold at the idling threshold w_0^0 of this no due-date problem and if in the original problem all of the due-date lead times satisfy

$$\rho_i \bar{f}_i \geq \sqrt{\frac{K + w_0^0 \xi_4^{\{1,\ldots,N\}}}{\xi_3^{\{1,\ldots,N\}}}} \left(\alpha_i \gamma_1 + \frac{\rho_i(1-\rho_i)}{2} \right) - w_0^0 \alpha_i \gamma_2, \tag{56}$$

then none of the products will ever be binding. Since the derivative of $x_i^{c*} - \tau\rho_i(1-\rho_i)/2$ with respect to total workload w is positive for the $s = 0$ case, we need only check that a product class is non-binding at the smallest workload experienced by the policy, which is the idling threshold w_0^0. Equation (56) guarantees this. Therefore, for due-date lead times satisfying (56), the optimal long run average cost remains constant for the same reason as stated in the standardized case.

Similarly, if region I conditions hold at the idling threshold w_0^0 and if the due-date leadtimes satisfy

$$\rho_i \bar{f}_i \geq \sqrt{\frac{K}{\xi_1^{\{1,\ldots,N\}}}} \left(\frac{\rho_i(1-\rho_i)}{b_i + h_i} \left[\frac{b_i - h_i}{2} + h_{\{1,\ldots,N\}} \right] + \frac{\rho_i(1-\rho_i)}{2} \right) \quad \text{for } i \neq \theta^* \tag{57}$$

and

$$\rho_{\theta^\bullet} f_{\theta^\bullet} \geq \sqrt{\frac{K}{\xi_1^{\{1,\ldots,N\}}}} \left(\frac{\rho_{\theta^\bullet}(1-\rho_{\theta^\bullet})}{2} + \sum_{\iota \neq \theta^\bullet} \frac{\rho_\iota(1-\rho_\iota)}{b_\iota + h_\iota} \left[\frac{b_\iota - h_\iota}{2} + h_{\{1,\ldots,N\}} \right] \right) - w_0^0. \quad (58)$$

then no product will be binding (one only needs to check at the idling threshold for the same reason as in the region II case).

For positive setup times, Property 3 states that as long as there is a customized product which does not have the cheapest tardiness cost, that product will eventually be binding for high total workload levels. The longer the due-date lead times the larger the total workload level before the product becomes binding. Although the average cycle cost is high for the workload levels where the product class is binding, the density associated with these regions diminishes rapidly. We can infer from this that as the due-date lead times increase, eventually the reduction in long run average cost asymptotically approaches the case where all of the products are treated as if they were standardized. Longer due-date lead times achieve diminishing returns for large total workloads.

5. DISCUSSION

In this section, we step through each region of the Venn diagram and discuss the relevant literature along with the insights derived from our analysis of the deterministic due-date lead time case; for ease of reference, the subsections below are numbered in the same way as the regions in Figure 1. Because our focus is on the dynamic stochastic versions of these problems, we do not include the vast literature on deterministic or static stochastic scheduling. For each subproblem, we compute and display a typical set of switching curves on the workload plane that characterize the proposed policy in the two-product case. The cost parameters satisfy $h_1 = 2h_2$ and $b_1 = 2b_2$ in each of these examples. Although our computations are restricted to the deterministic due-date lead time case, Markowitz undertakes a partial examination of the cost per cycle for a customized product with uniform due-date lead times, and the numerical results suggest that the qualitative observations described in this section carry over to the random due-date lead time case.

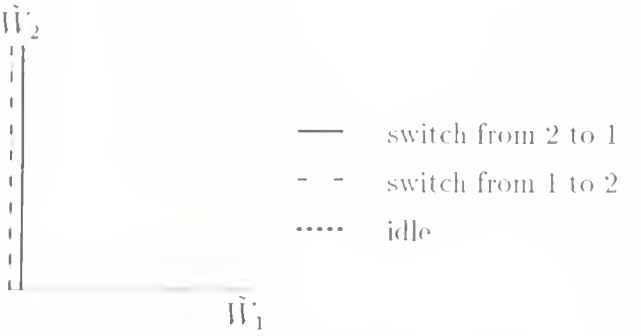

Figure 3: Customized products, no setups, no due-dates.

5.1. Customized Products, No Setups, No Due-dates. The outer area of Figure 1 corresponds to systems with only customized goods, no setup penalties and no due-dates. These traditional multiclass queues are the simplest of the systems displayed in the diagram and have been extensively studied. Cox and Smith (1961) were the first to show the optimality of the "$c\mu$ rule" (in our notation, the $c\mu$ index corresponds to the tardiness cost rate b), which gives priority to the product that, while serviced, removes cost from the system at the highest rate; see Bertsimas and Niño-Mora (1996) for an up-to-date set of references for this scheduling problem.

Our proposed policy is similar in spirit. The policy parameters can be easily determined from §4.2. The lack of setups forces region II to vanish and the cycle length τ to be zero. No due-dates indicates that $f_i = 0$ for all products, implying that only the region III conditions can apply. Thus for any total workload $w > 0$, the cycle center x_i^* is set to zero for all i not equal to θ_b^*, and the cycle center for the cheapest backorder product is set to w. In this special case, one can trivially calculate that the idling threshold w_0 is zero. The implied dynamic cyclic policy is then simple in form (see Figure 3 for a two-product example where $b_1 > b_2$): Service all but the least expensive product to exhaustion and switch out of producing the cheapest product if there are any higher cost products present. This policy can be interpreted as a two-level priority rule: All but the least expensive product have high priority and are served to exhaustion in a cyclic manner; the least expensive product has low priority.

It is worth noting that in the heavy traffic limit the queue length of the high priority products vanish and only the lowest priority product is present (see Whitt 1971). It follows that the heavy traffic limiting behavior of the proposed policy is identical to that of a strict "b" priority policy: In both cases, the workload process lies along the θ_b^* axis. Thus, in the heavy traffic setting nothing is lost between our proposed dynamic cyclic policy and the $c\mu$ rule.

5.2 Standardized Products, No Setups, No Due-dates. Multiclass queueing systems with standardized products are generally considered to be more difficult to analyze than systems with customized products because of the nonlinear cost structure introduced by having both holding and backorder costs. Additionally, when constructing policies for these production/inventory systems, there are no natural switching boundaries as with exhausting a queue in customized systems. These obstacles have hindered the analysis of even the simplest case: Standardized products with no setups and no due-dates. Zheng and Zipkin (1990) analyze the intuitively appealing longest queue (or smallest finished goods inventory) policy for a two-product system, which is optimal when both products have identical cost parameters and operate under independent base stock policies. Ha (1993) partially characterizes (in terms of switching curves) the optimal policy for the Markovian two-product case. Wein (1992) uses heavy traffic theory to develop a dynamic priority policy and an aggregate base stock policy for the multiproduct problem, and Veatch and Wein (1993) expand upon this by examining index policies in a Markovian setting and by analyzing the lost sales case. Finally, Peña and Zipkin (1993) propose a policy that combines the best aspects of the policies in Zheng and Zipkin, Ha, and Wein.

Our proposed policy is similar to that of Wein, and readers are referred there for a more in-depth discussion of the basic insights gained from the heavy traffic analysis of this system. From the calculations involving setup penalties and cycle length in §4.2, τ is again zero for this case and only regions I and III are possible. The cycle center of all but the cheapest product is set to zero and $x^*_{\theta_b}$ equals w for $w > 0$ and $x^*_{\theta_b}$ equals w for $w < 0$. Because there are no setup times, the total workload process, which measures the total work in backorders minus the total work in finished goods inventory, is a RBM and

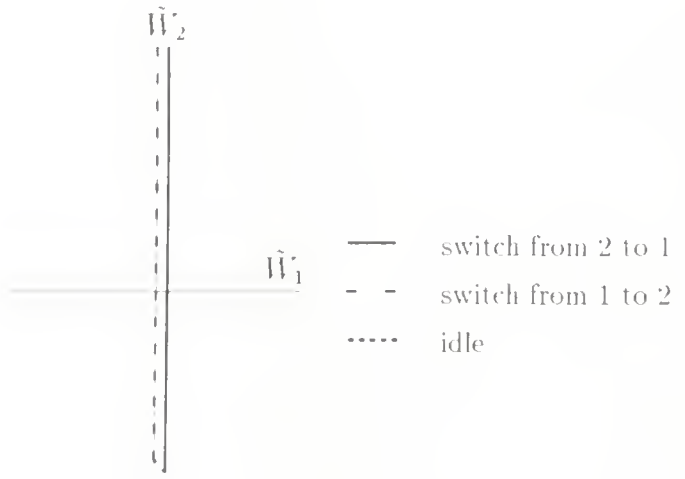

Figure 4: Standardized products, no setups, no due-dates.

the idling threshold w_0 is negative (i.e., $-w_0$ represents an aggregate base stock level for the total work in finished goods inventory) and has an explicit form (see equation (76) of Wein). The dynamic cyclic policy can be stated as a two-level priority system for unfilled orders with the added complication of a mechanism to build up a finished goods inventory. It has the following rules: 1) service all orders with finished goods inventory if available; 2) all queued orders for products other than product θ_b^* have priority and are serviced in an exhaustive cyclic manner; 3) orders for the cheapest backorder product θ_b^* have the lowest priority; 4) if no orders are present and the total workload is above the idling threshold, then produce product θ_h^*, which is the cheapest to hold in inventory. A typical two-product policy (where $b_1 > b_2$, $h_1 > h_2$) is pictured in Figure 4 and, as in the customized case, the strict priority rule between the products causes the switching curves to nearly overlap on product 2's axis. The performance of the resulting policy is, in the heavy traffic limit, identical to the one suggested by Wein.

Comparing the customized product and standardized product cases, we see from the proposed policy a distinct role for inventory: It hedges against the risk of backordering. According to the HTAP, uncertainty in production and demand have their greatest impact on the total workload in the system, and the time scale decomposition allows the

machine to flexibly address requests for high cost products without the need for storing the products themselves. Thus, inventory acts as a reservoir of reserve capacity, able to absorb random fluctuations in service and demand, and the proposed policy stores this capacity in the most economical manner possible: It is placed in the cheapest holding cost product. Similarly, when the finished goods inventory is exhausted, the proposed policy is still able to flexibly service requests for high cost products by neglecting the cheapest backorder product, effectively storing deficit inventory in its cheapest form.

5.3. Customized Products, Setups, No Due-dates. The models represented by region 3 in the Venn diagram are known as polling systems in the literature on computer communication networks. Although considerable research has appeared on the performance analysis of these systems (e.g., Boxma and Takagi, 1992), the dynamic scheduling of these multiclass queues with setups have not yielded to an exact analysis. Hofri and Ross (1987), Liu, Nain and Towsley (1992) and Koole (1994) derive structural results. Reiman and Wein (1994) use heavy traffic analysis to develop policies for the two-product problem. Boxma, Levy and Westrate (1994) and Bertsimas and Xu (1993) compute static policies (i.e., polling tables), Browne and Yechiali (1989) derive a quasi-dynamic index policy to choose sequences of products to service at the start of each cycle, and Duenyas and Van Oyen (1995, 1996) develop scheduling heuristics based on myopic reward rates for systems with setup times and costs, respectively; readers are referred to the literature review of Reiman and Wein (1994) for more details.

Our proposed policy is a multiproduct version of Reiman and Wein's (1994) policy. As in the previous customized case, the lack of due-dates and of standard goods limits the cycle center and cycle length equations to the region III formulations. Given that f_i equals zero for all i, all but the cheapest product, θ_b^-, is binding and so $\Theta^* \subset \{\theta_b^-\}$ for all S and all w. Thus, the policy can be stated as follows: Service each product in a cyclic manner. When set up for all but the cheapest backorder product, service it to exhaustion. When set up for θ_b^- work on it until its normalized workload reaches $x_{\theta_b^-}^c(w) - \tau(w)\rho_{\theta_b^-}(1-\rho_{\theta_b^-})/2$, where w is the current normalized work in the system. When there are only two products, the policy is identical to Reiman and Wein's (1994).

The presence of setups eliminates the two-level priority scheme seen in the previous

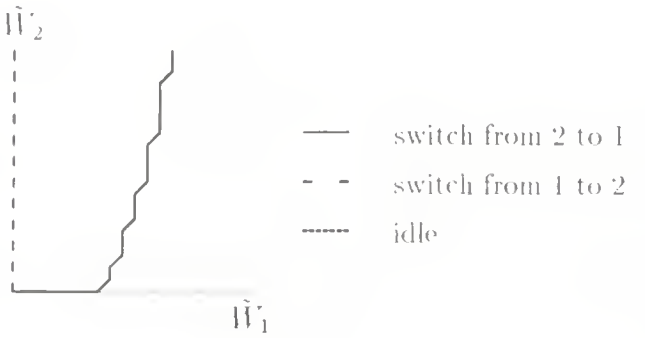

Figure 5: Customized products, setups, no due-dates.

no setup cases, because a strict priority rule leads to excessive setups. The proposed policy avoids this by keeping to the cycle, yet minimizes cost by offering quick delivery to high cost products at the possible neglect of the θ_b^* product. However, as in the previous two subproblems, the heavy traffic analysis essentially treats all but the lowest priority product in a similar fashion. These results suggest that for heavily loaded polling systems, it is more beneficial to focus on noncyclic exhaustive policies than dynamic non-exhaustive policies.

A typical two-product policy is shown in Figure 5. The presence of setup penalties has added breadth to the switching curves of Figure 3. The policy can be viewed as not only balancing setups and queueing costs, as directly seen in the formulation of τ and x^c, but also as controlling the randomness of the system by isolating the effects of total work fluctuation in the least cost product. This is seen in Figure 5 by the huge range of low cost product queue length values (along the vertical axis) caused by the fluctuation of total workload in the system versus the relatively confined range of the more expensive product (along the horizontal axis).

5.4. Standardized Products, Setups, No Due-dates. Standardized goods with setups and without due-dates has long been considered the prototype for modeling make-to-stock manufacturing systems. The deterministic version of this problem, which is called the Economic Lot Scheduling Problem (ELSP), was originally formulated in 1915 (see Elmaghraby 1978), and only recently has the stochastic version of this problem

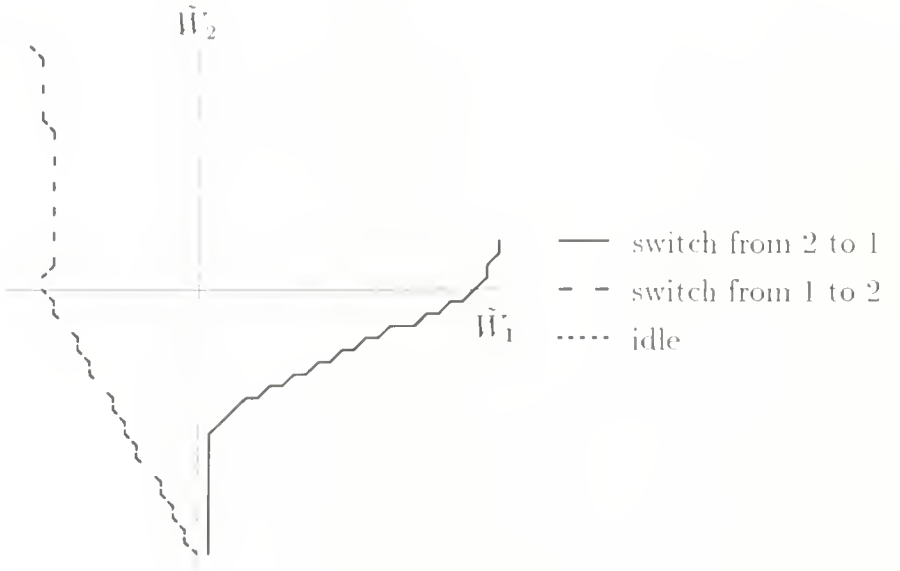

Figure 6: Standardized products, setups, no due-dates.

(the SELSP) received attention. Graves (1980) develops a heuristic using a Markov decision model. Leachman and Gascon (1988), Gallego (1990) and Bourland and Yano (1995) develop heuristic lot-sizing algorithms that are rooted in the deterministic ELSP. Sharifnia, Caramanis and Gershwin (1991) use a hierarchical decomposition approach to analyze a stochastic fluid version of the problem, Federgruen and Katalan (1995b, 1996) use polling theory to analyze the performance of (periodic and cyclic, respectively) base stock policies for the SELSP, Qiu and Loulou (1995) numerically compute the optimal solution to the two-product problem by modeling it as a semi-Markov decision process, Anupindi and Tayur (1994) develop a simulation based approach to compute cost-effective base stock policies, MRW analyze the problem using the HTAP, and Sox and Muckstadt (1995) formulate the problem as a stochastic program, and propose a decomposition procedure to solve it.

Our proposed policy restricted to standardized goods and no due-dates reduces to that of MRW. Since all of the products are standardized, there are no orthant constraints and hence no binding products; i.e., $\Theta^* = \{1, \ldots, N\}$. All three regions are possible and the cycle length, cycle center and idling threshold are nontrivial. A sample policy is

pictured in Figure 6.

A detailed description of the proposed policy for the SELSP and a summary of the key insights are given in MRW. Here we briefly cover the highlights as they pertain to the larger problem. As in the standardized goods, no setup case, the policy treats the total workload inventory as a reservoir of stored capacity, used to hedge against demand and service rate uncertainty. The switching curves are constructed so that excess inventory is stored in the cheapest holding cost product θ_h^*, and deficit inventory is moved into the cheapest backorder product θ_b^*. Positive inventory also serves a secondary role: A small cache of inventory impedes backordering on an individual product level while the machine is producing other products during a cycle. The amount of inventory required is dependent on the length of the cycle, which is an increasing function of the setup penalty as shown in equation (51). Also, setups introduce breadth between the switching curves, as seen in comparing Figures 4 and 6.

5.5 Customized Products, No Setups, Due-dates. The systems in region 5 of Figure 1 reflect manufacturing facilities that service customer requests on a make-to-order basis with the additional feature that customers do not want the goods immediately but at some future time. The inclusion of due-dates causes an explosion in the dimension of the state space, which makes this problem difficult. Pandelis and Teneketzis consider earliness and tardiness penalties and examine properties of an optimal policy. Righter uses stochastic ordering to further characterize aspects of an optimal policy. Van Mieghem (1995) studies a multiclass queueing system with costs based upon a convex nondecreasing function of each order's delay in the system. Using heavy traffic analysis, he shows that a *generalized* $c\mu$ rule is asymptotically optimal, where c is a dynamic function that represents each job's marginal cost of delay.

As in both no setup, no due-date cases, the lack of setup penalty causes the cycle length τ and region II to vanish. The presence of (deterministic) due-dates, however, moves the switching curves into the orthant so that they lie on a new set of axes; i.e., by equation (40) the cycle centers x_i^c are shifted by $\rho_i f_i$. The proposed policy can be described as follows: If the total work in unfinished orders is less than $\sum_{i=1}^{N} \rho_i f_i$, then orders that are almost at their due-date have priority and are serviced in a cyclic manner.

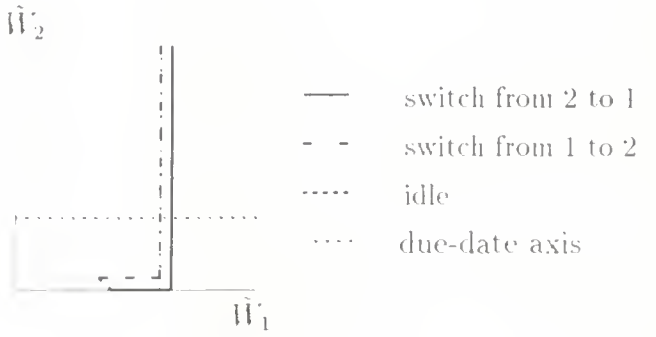

Figure 7: Customized products, no setups, due-dates.

and if there are no orders near their due-date and the total workload level is above the idling threshold, then the machine works on orders for the product with the smallest earliness cost. If the total workload is greater than $\sum_{i=1}^{N} \rho_i f_i$ then all products other than the cheapest tardiness product θ_b^* have priority and their orders are serviced in a cyclic manner just as their due-dates are reached or passed; the cheapest tardiness product θ_b^* has lowest priority and its orders are serviced only when the due-dates of the higher priority goods are distant.

A two-product example is given in Figure 7. As in the no due-date, no setup examples, the switching curves for the two products nearly overlap. However, the switching curve shift onto the new due-date axes is readily transparent by comparing Figures 4 and 7. The shift of the due-date axes in the workload plane represents the tolerance of the policy toward aging orders not due in the near future. The region corresponding to the "northeast" portion of the plane corresponds to the workload states where there is too much work and orders are completed past their due-date. The "southwest" portion of the plane is an area where orders are few and if worked on will be completed early. The intersection of the new axes (the vertical one is hidden behind the switching curve) corresponds to the state where the wait in queue for an order exactly equals its due-date lead-time.

As in the previous subproblems, the proposed policy minimizes the costs of the higher cost products at the expense of the cheapest product. The policy attempts to service the

high cost products in such a manner that they are completed exactly when they are due. As in the no due-date case, excess orders are shifted to the cheapest tardiness product θ_b^*. The presence of due-dates allows the scheduler to avoid tardiness costs by staying "ahead of schedule" - that is, by completing some of the orders early to allow for more slack when there is an unexpected surge in demand or difficulty in production. It hedges against this uncertainty in the most economical manner possible: It only completes early those products with the cheapest earliness cost. However, customized products have a limit on how much "deficit" workload they can hold: The policy is forced to switch setup when a product is exhausted of orders, and, as in the "kink" on the horizontal orthant boundary in the proposed switching curves in Figure 7, further hedging against tardiness must be achieved by completing early the more expensive product. In addition, given that the low total workload is costly, the idling threshold may be nonzero so as to avoid states with high earliness costs.

It is interesting to compare these results to those of Van Mieghem. His cost structure can accommodate a multiclass queueing system with customized products, no setups and deterministic due-dates, where the earliness costs h_i are set to zero for all products. In this case, Van Mieghem's generalized $c\mu$ rule gives priority to tardy orders over early orders (and early ones can be processed in any manner), and priority within the tardy group is determined, in out notation, by the highest "b" rule. Also, the server works as long as there are orders waiting.

Our results provide the dynamic cyclic version of this policy just as was seen in the standardized and customized cases with no due-dates and no setups. There is a two-level priority scheme for tardy orders, where all but the cheapest tardiness cost product θ_b^* have priority and are serviced cyclically; tardy θ_b^* orders are serviced only when there are no other orders past due. Since all of the earliness costs are equal, the proposed policy services the orders to exhaustion when there are no tardy orders. The idling threshold is set at zero and so does not contribute to the policy: The machine works as long as there is work to do. Again, as in the previous cases without setups, the dynamic cyclic policy has the same behavior as the generalized $c\mu$ rule in the heavy traffic limit. Thus, our policy and Van Mieghem's are similar when they are both restricted to the one case

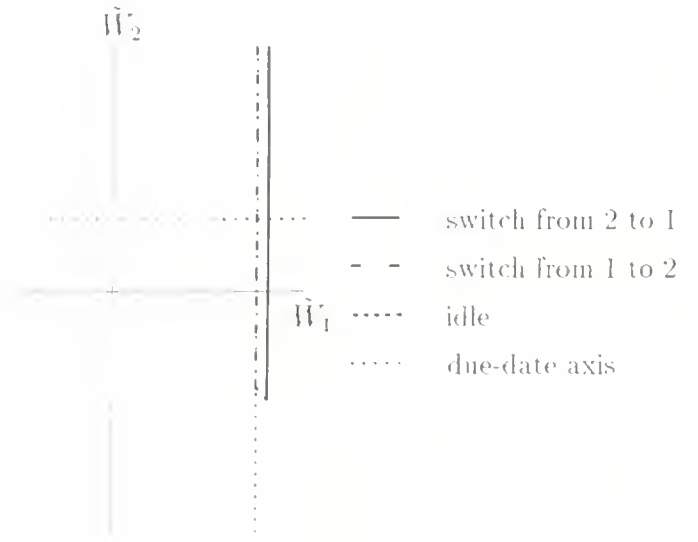

Figure 8: Standardized products, no setups, due-dates.

where the models overlap.

5.6. Standardized Products, No Setups, Due-dates. This case, which has not been explicitly studied to our knowledge, is similar to the customized case, but without the orthant boundaries. The cycle length and region II vanish. The cycle centers are shifted onto the new due-date axes $\rho_i f_i$. The presence of a finished goods inventory, however, modifies the proposed policy from the previous customized one. The policy can be interpreted as the following rules: 1) only fill orders when they are due, either from the finished goods inventory or directly from the machine output; 2) if the total workload present in orders minus that in finished goods inventory is above $\sum_i \rho_i f_i$ then follow a two-level priority scheme: Orders for high tardiness cost goods (i.e., all but θ_b^*) that are closer to their due-date have priority and are serviced in a cyclic manner and tardy θ_b^* orders have lower priority; 3) if total work is below $\sum_i \rho_i f_i$ then tardy orders have priority and are serviced cyclically, and if there are no tardy orders the machine works on the lowest holding cost product θ_h^* and stores them in the finished goods inventory, until the workload idling threshold is reached. A typical two-product policy is depicted in Figure 8.

This policy is exactly the same as a *shifted* standardized, no setup, no due-date

policy. Moreover, in the heavy traffic limit *both systems incur exactly the same long run average cost*. This perhaps surprising result makes intuitive sense upon closer inspection. The due-dates we have considered have a special structure: They are $O(\sqrt{n})$ and so only influence the fluid limit. This implies that in our policy, orders will arrive, become late and be serviced before the total workload has an opportunity to significantly change. Moreover, the orders are continuously arriving and in this time frame the machine is not able to either get ahead or fall behind on orders. Thus, if we are servicing orders that are due today and arrived 10 days ago, then tomorrow we will be servicing orders due tomorrow that arrived nine days ago. Due-date lead times have not provided any additional flexibility to the system; orders are not serviced earlier or later than usual, only the absolute time of service has been shifted.

The difference between this case and the customized one is the ability to pre-make goods for a finished goods inventory. With standardized goods, the policy is always allowed to hedge against backordering by investing in stored work in its cheapest form. In the customized case (see Figure 7), the inability to manufacture goods in anticipation of future orders means that the policy might be forced to store work in a more expensive product when it runs out of orders for θ_h^* products and so must finish early the next cheapest holding cost product.

In a system not in heavy traffic, however, the long run average cost is affected by the due-date lead time. Buzacott and Shanthikmar (1993, §4.5) find the long run average cost for the single-product Markovian version of this problem. If one optimizes over the idling threshold (Buzacott and Shanthikumar's target level), then the long run average cost is

$$\frac{h}{\ln \rho} \ln \left(\frac{h}{b+h} \right) + h \tilde{f} \left(\mu \frac{1-\rho}{\ln \rho} + \lambda \right), \tag{59}$$

as long as the due-date lead time $\tilde{f}$ is less than or equal to $\frac{1}{\mu(1-\rho)} \ln \left(\frac{h+b}{h} \right) - w_0$ (this condition guarantees that the idling threshold w_0 is nonpositive). The first term in (59) is equal to the long run average cost for a system without due-dates and the second is the due-date contribution. Since $\mu \frac{1-\rho}{\ln \rho} + \lambda$ is always negative, due-date lead times reduce the cost of the system.

The reason for the difference between the heavy traffic system and the unscaled one can be most easily understood by examining the unscaled system with no orders and a full inventory. When an order arrives, the total order workload increases above the idling threshold and the server initiates production. Goods can be finished before the order is due, inflating the inventory level beyond the heavy traffic prediction. This has two effects: Holding costs are greater and the chance of backordering is decreased. The due-date leadtime has allowed the system to hedge against backordering, something that the heavy traffic approximation has not allowed for. Nonetheless, as utilization grows close to one, the long run average cost in (59) is dominated by $\frac{h}{\ln \rho} \ln \frac{h}{b+h}$, which is independent of the due-date lead time.

It is also interesting to note the difference between our system and the inventory model of Hariharan and Zipkin (1995). They consider a facility that employs a one-for-one replenishment scheme for a single product, and optimize the on-site inventory level. There is a deterministic due-date lead time, L_d, for requests and a lead time for the facility's orders, L_s. They find that increasing the due-date lead time decreases the average cost of the system up until the due-date lead time equals the re-order lead time; when $L_d \geq L_s$ further increases in the due-date lead time have no value. One might be tempted to conclude that the two results are consistent, because $L_d \geq L_s$ for our system. However, in our problem, the sojourn time for orders plays the role of L_s, and it is also $O(\sqrt{n})$ by the heavy traffic conditions; hence, $L_d \geq L_s$ does not always hold in our system. We believe the reason for the difference between Hariharan and Zipkin's result and ours is that Hariharan and Zipkin consider an uncapacitated system, where inventory hedges against demand uncertainty but does not act as stored workload, freeing machine resources for higher cost products. With longer due-date lead times there is less urgency for orders to be replenished quickly and so less inventory is needed on-site to hedge against backordering. In contrast, in our capacitated model, the same amount of stored work is needed independent of due-date lead time, as discussed above.

5.7. Customized Products, Setups, Due-dates. We know of no previous work analytically treating this problem. This case contains all of the complexity of §4.2. All three regions can be present, cycle length, cycle center and idling threshold are nontrivial

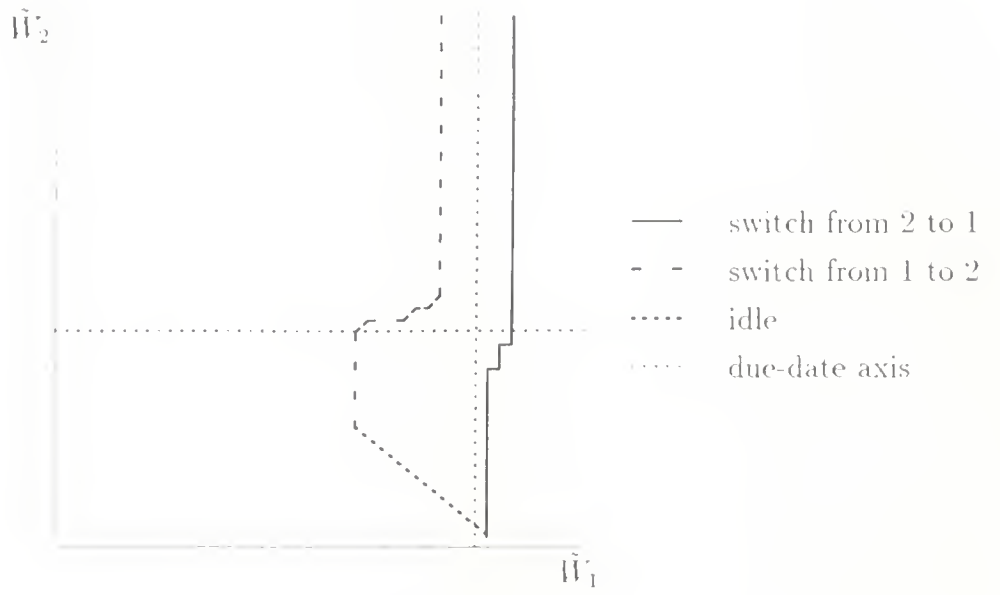

Figure 9: Customized products, setups, long due-dates.

and products can be binding. It is interesting to note that if the due-date lead times are long enough (Figure 9), the proposed policy looks like a shifted set of SELSP switching curves (systems with setup times, however, will always be slightly different because the expanding cycle length τ will eventually hit the orthant boundary for large total workload, by Property 1 in §4.4). If the due-dates are short (Figure 10), the switching curves bump into the orthant and flatten out. Readers should note that Figures 9 and 10 are based on cases with setup costs but no setup times, whereas the other cases with setups in this section contain setup times but no setup costs. In addition to viewing the case as a shifted SELSP policy, it can also be thought of as the customized product, no setups, due-dates case (Figure 7) with breadth added to the switching curves as was seen in the transformation between the no due-date cases without setups to the case with setups (Figures 3 and 5).

5.8. Standardized Products, Setups, Due-dates. This subproblem corresponds to the center of Figure 1, and has received very little attention. The switch to standardized products from the customized case eliminates the orthant boundaries, thereby simplifying the nature of the optimal dynamic cyclic policy. The policy is merely the SELSP

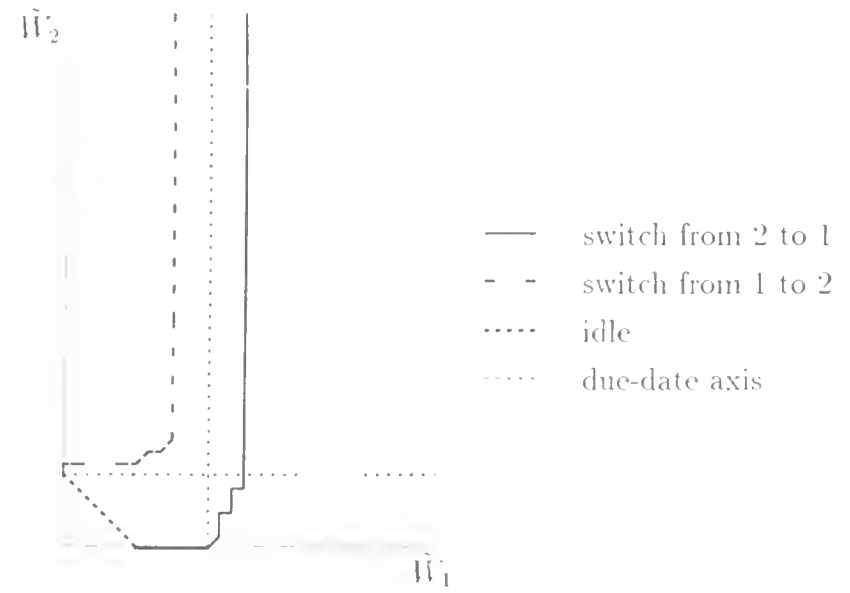

Figure 10: Customized products, setups, short due-dates.

policy on the shifted set of due-date axes (see the example in Figure 11 and compare it to Figure 6); moreover, as in §5.6, the cost under the optimal dynamic cyclic policy is insensitive to due-date lead times. Markowitz performs simulation runs for the $\rho = 0.9$ case that support this insensitivity conjecture.

It is worth noting that as due-date lead times increase, the entire set of switching curves passes into the positive orthant. Hence, longer due-date lead times can transform a make-to-stock method of servicing demand for standardized goods into a make-to-order method. When due-date lead times reach this critical level, it is optimal to fill orders early and incur holding costs; these early orders provide a buffer against backordering as would a finished goods inventory.

5.9 Systems with Customized and Standardized Products. Lastly, we consider *mixed* systems with both customized and standardized products, which would correspond to the border of the "Standardized" circle in Figure 1. Our distinction between customized and standardized is based entirely on product design, and the derived switching curves dictate which standardized products are made-to-order. In contrast, other work has treated the partition of customized and standardized goods as a decision to be

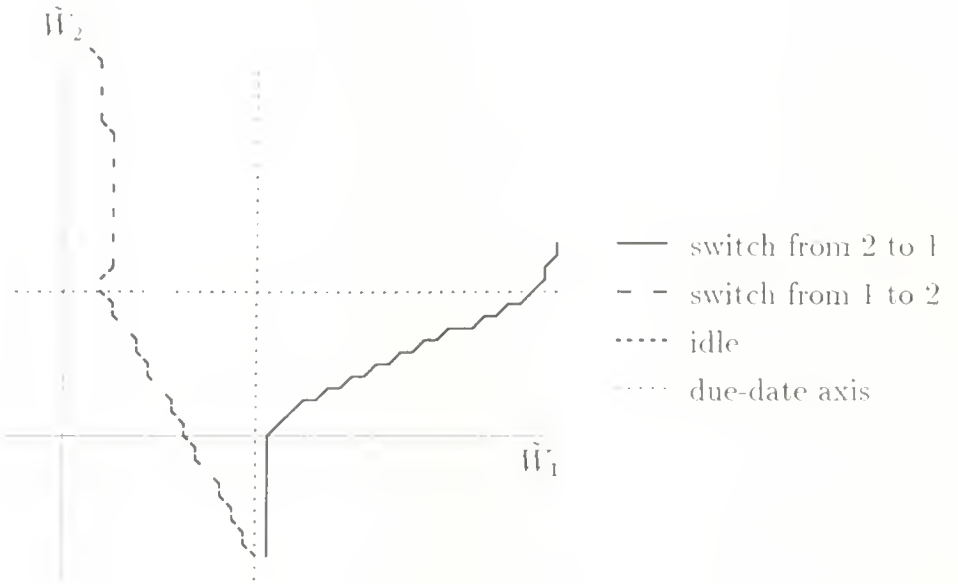

Figure 11: Standardized products, setups, due-dates.

optimized. For example, Carr *et al.* (1993) consider a queueing system with no setups or due-dates, where the make-to-order (MTO) goods represent low demand items that have priority over the make-to-stock (MTS) ones. This intrinsic priority rule allows for a performance analysis of the MTO/MTS partition, but does not involve optimal scheduling of the products. Nguyen (1995a) performs a heavy traffic analysis of mixed systems without setups or due-dates, and with lost sales instead of backordering. In a subsequent paper (1995b), she examines different priority rules for the MTO and MTS products and suggests an algorithm for setting base stock levels. Federgruen and Katalan (1995a) examine mixed systems with setups and no due-dates, and compare several priority rules for switching from MTS goods to MTO. They also propose a heuristic for partitioning MTO and MTS items.

Finally, there is the full problem: Customized and standardized products, setups and due-dates. All of the previous cases are subsets of this general model. It is the proper setting to ask questions of balancing inventory costs and setup penalties, of setting base stock inventory levels and avoiding backordering, of determining due-date lead time effects and natural due-date-based partitions of MTS/MTO goods. Yet, this generality

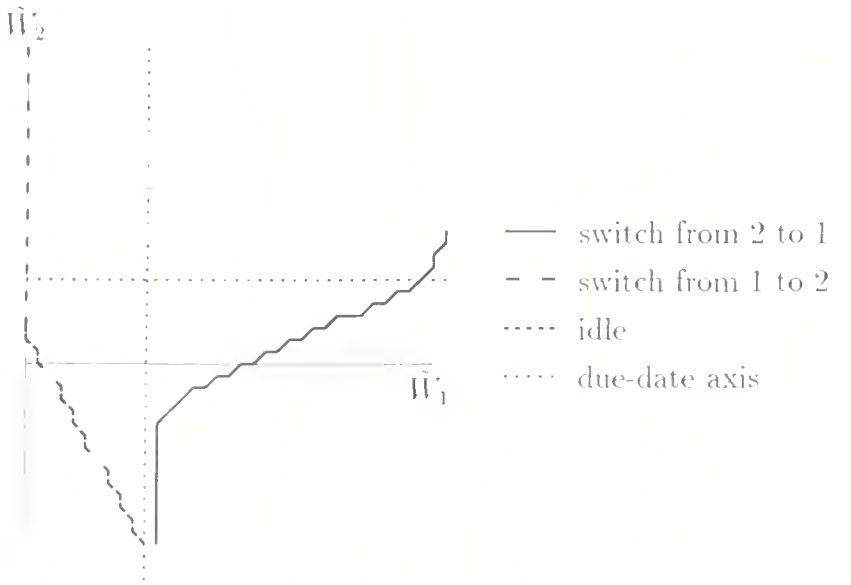

Figure 12: Mixed products, setups, due-dates.

does not complicate the system beyond the previous cases. The product mix is accounted for in our calculations by the presence of orthant constraints on some products and not on others (see Figure 12 for an example of a mixed system with one customized product, one standardized product, setup times and due-dates). Most of the insights described earlier about the interactions of setups, due-dates and orthant constraints carry over directly to the combined system.

6. COMPUTATIONAL STUDY

Computational results in Reiman and Wein (1994) and MRW confirm the accuracy and robustness of the HTAP for make-to-order and make-to-stock systems, respectively, with either setup costs or setup times, but no due-dates. Here we report on a computational study that tests our methods with respect to mixed systems and due-dates. Unfortunately, the presence of due-dates prevents an exact derivation (via dynamic programming) of the optimal policy and an exact evaluation of the various policies under consideration. Therefore, we employ discrete event simulation for these tasks. For sim-

plicity, we only examine two-product systems with deterministic due-date lead times. Interarrival times, service times and setup times are independent and exponentially distributed. The service is preemptive-resume and there are no setup costs. For all of the scenarios in this section, we start with systems void of both orders and inventory and then perform ten independent runs. Each run contains a 100,000 time unit initialization period and statistics are recorded for the next 10,000,000 time units.

We wish to examine three issues for mixed systems: The accuracy of the heavy traffic approximation, the effectiveness of our proposed policy, and the value of increased due-date lead times. The first product is customized and the second is standardized. Using the same parameters as a case in MRW, we set the earliness costs to $h_1 = 2, h_2 = 1$, tardiness costs to $b_i = 5h_i$, service rates $\mu_i = 1$, arrival rates $\lambda_1 = 0.6, \lambda_2 = 0.3$ and average setup time per cycle $s = 20$. We set the due-date lead times $f_1 = f_2$ and test f_i for values equal to 0, 20 and 100.

Straw Policies. Since we do not have a convenient point of reference provided by an optimal policy, we compare the proposed policy, which is defined in equations (26)-(27), to three straw policies. The first straw policy is the proposed policy without due-date considerations; that is, the policy in (26)-(27) is calculated with the due-date lead time f_i set to zero. The second straw policy is a hybrid base stock/exhaustive policy where the customized product is serviced to exhaustion and the standardized product is produced up to a base stock level $\mu^{-1}\tilde{v}$, which is equivalent to a $-\tilde{v}$ order workload level. The base stock level is calculated in a fashion analogous to the standardized, no due-date case in §3.1 of MRW, as follows. The cycle center for the customized product (product 1) is set to $\tau(w)\rho_1(1-\rho_1)/2$. The cycle center for the standardized product is set to $\tau(w)\rho_2(1-\rho_2)/2 + \tilde{v}$. If we define $v = \tilde{v}/\sqrt{n}$, then the cycle length is $\tau(w) = 2(w-v)/[\rho_1(1-\rho_1)+\rho_2(1-\rho_2)]$. Under this policy, the total normalized workload W has a stationary gamma density with parameters $\alpha = 2\sqrt{n}(1-\rho)/\sigma^2$ and $\beta = s\sum_{i=1}^{2}\rho_i(1-\rho_i)/\sigma^2$ (see Coffman, Puhalskii and Reiman 1995b). The average cost for this policy is

given by

$$\int_{-v}^{\infty} \left(c_1\left(\frac{w-v}{\rho_1(1-\rho_1)\sum_{i=1}^{2}\rho_i(1-\rho_i)}, 2\frac{w-v}{\sum_{i=1}^{2}\rho_i(1-\rho_i)}, w\right) \right.$$
$$\left. + c_2\left(v + \frac{w-v}{\rho_2(1-\rho_2)\sum_{i=1}^{2}\rho_i(1-\rho_i)}, 2\frac{w-v}{\sum_{i=1}^{2}\rho_i(1-\rho_i)}, w\right) \right) \frac{\alpha(\alpha(w-v))^{\beta}}{\Gamma(\beta+1)} e^{-\alpha(w-v)} dw. \tag{60}$$

We set n equal to $(1-\rho)^{-2}$ and use a steepest descent algorithm to find the parameter v which minimizes (60).

The third straw policy is again a hybrid base stock/exhaustive policy where the base stock level is determined by an exhaustive search using multiple simulation runs (the results of which are not shown here). These two hybrid policies, which will be referred to as the *hybrid HTAP* policy and the *hybrid search* policy, together offer the opportunity to determine the accuracy of the HTAP. Numerical results are contained in Table 1 and switching curves for the hybrid search policy and the proposed policy are given in Figures 13 and 14 for the $f_i = 0$ and $f_i = 100$ cases, respectively.

Due-Date Lead Time	Cost of Hybrid HTAP	Cost of Hybrid Search	Cost of Proposed w/o D-date	Cost of Proposed Policy
0	290.3 (± 3.2)	290.0 (± 2.7)	274.6 (± 1.9)	274.6 (± 1.9)
20	201.8 (± 2.7)	199.5 (± 3.2)	186.3 (± 2.1)	184.1 (± 2.7)
100	122.8 (± 1.4)	121.3 (± 1.3)	158.3 (± 0.9)	109.7 (± 1.1)

Table 1: Simulation results for the mixed system.

Observations: Straw Policies. The long run average cost for both hybrid policies decreases with longer due-date lead times. With zero due-date lead times, orders for the high cost customized product are immediately backordered, driving up costs. As the due-date lead times increase, orders for the customized product are less tardy and costs fall. However, one would expect that eventually they would start to increase again (under the hybrid policies) as the due-date lead times become inordinately long and holding costs become excessive.

The two hybrid policies incur nearly identical costs. The base stock levels for the search policy are 45, 37 and 13 for the $f_i = 0$, 20 and 100 cases, respectively, and the corresponding HTAP levels are 39, 28 and 5. Although the base stock levels for the

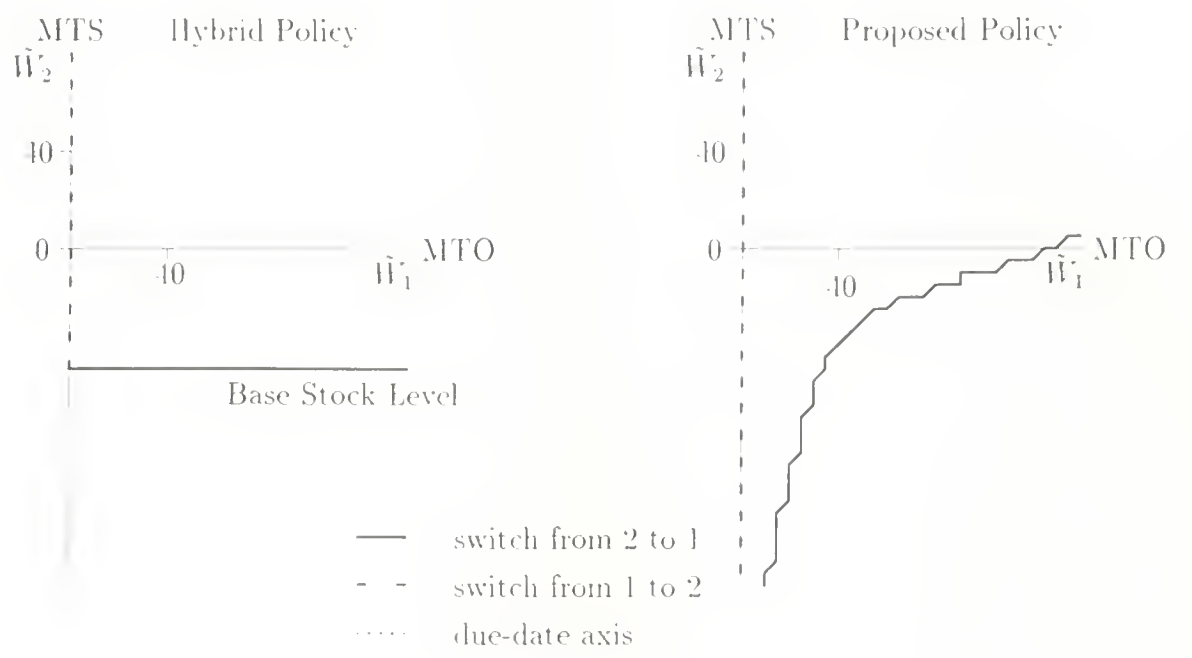

Figure 13: Switching curves for a mixed system with zero due-date lead times.

search policy are consistently higher than those of the HTAP policy, the actual difference in long run average cost is very small. For the hybrid exhaustive/base stock policy, the long run average cost as a function of base stock level appears to have a shallow slope about the optimal solution, and the heavy traffic analysis is able to identify a base stock level that performs quite well.

Finally, it is not surprising that the performance of the proposed policy that assumes zero due-date lead times deteriorates as due-date lead times increase. This deterioration suggests that it is not advisable to ignore due-dates when due-date lead times are large.

Observations: Proposed Policy. For the reasons cited above, the proposed policy also has decreasing long run average costs as due-date lead times increase. The proposed policy has an important advantage over the hybrid policies that can be seen in Figures 13-14: It is able to avoid large buildups of product 1 orders, both in queue and waiting to be shipped. The cost reduction achieved in Table I by the proposed policy relative to the hybrid policies is in the 5-10% range, and increases with the due-date lead time f_1. For the $f_1 = 0$ case, the proposed policy avoids severe backordering by switching to product

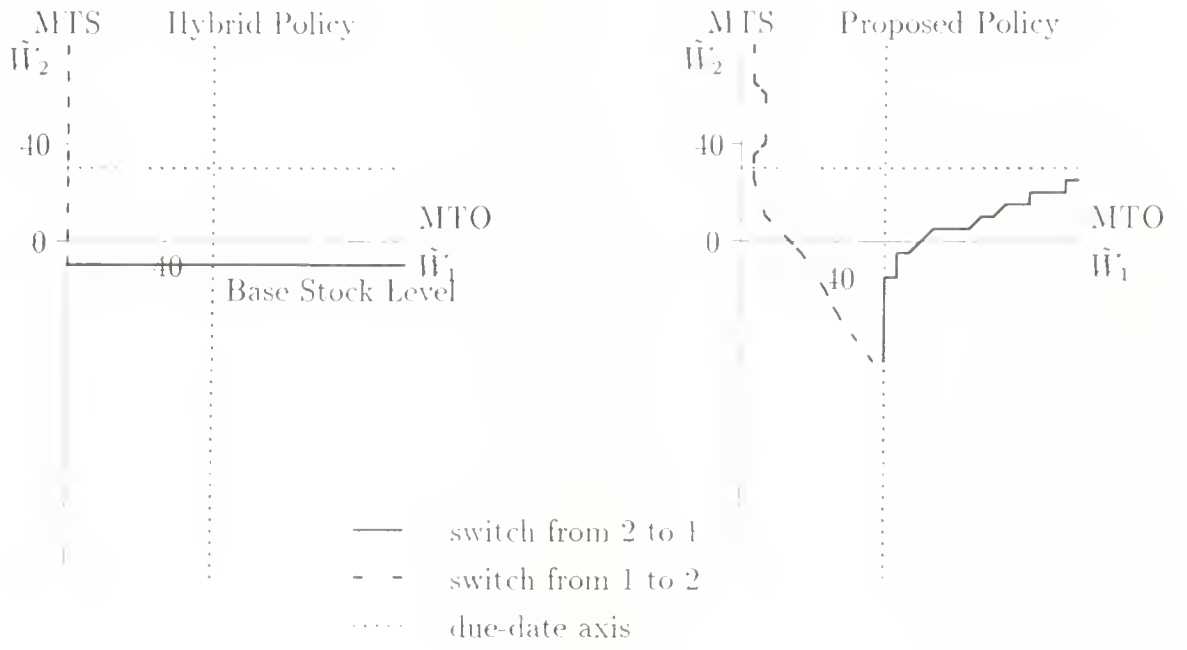

Figure 14: Switching curves for a mixed system with due-date lead times of 100.

1 if there is an excessive number of orders in queue. However, due to the severity of the setup penalty, the cycle length τ is long and so the product 1 buildup must be large. Thus, the marginal benefit of the proposed policy over the hybrid policy is not great in this case. As the due-date lead time increases, however, the proposed policy has a greater opportunity to avoid excessive product 1 costs. The cycle length τ is still large but the policy is able to attain a balance between earliness and tardiness costs by cycle center placement. As seen in Figure 14, the amount of product 1 workload is maintained near the due-date axis $\rho_1 f_1$, which is the level of product 1 workload necessary for a new order to wait in queue for an amount of time exactly equal to its due-date lead time. The hybrid policies are not able to perform this type of cost minimization.

The Value of Due-date Lead Times. As due-date lead times increase, our heavy traffic analysis (see §4.6) predicts that the long run average costs under the mixed system should decrease until the due-date lead times hit a critical value; beyond this threshold value, costs should be very insensitive to due-date lead times and the total system cost should be almost equal to the cost under the corresponding SELSP. Moreover, because the hybrid

policies optimize the standardized and customized products independently, the inventory costs for the standardized product should be independent of the due-date lead time under these policies.

For the hybrid policies, there is little change in the costs due to the standardized product in Table I: Its average holding and backorder costs remain approximately 23 and 14, respectively, for all values of the due-date lead time, and almost all of the savings are due to the customized product. In contrast, the inventory costs for the standardized product in the proposed policy change with the due-date lead time (in the $\tilde{f}_i = 20$ case the earliness cost is 16.7 and the tardiness cost is 12.7; in the $\tilde{f}_i = 100$ case, they are 34.7 and 16.2, respectively). This difference is due to the dynamic nature of the proposed policy. Unlike the hybrid policies, the proposed policy is able to shift inventory costs between the two products by changing the cycle center and cycle length for each workload level.

To analyze the costs incurred by the customized products, we compare the mixed system to its corresponding SELSP, which is MRW's asymmetric, setup time, $b_1 = 10$ case. According to the dynamic programming results in row 13 of Table VIII of MRW, the SELSP cost under the dynamic cyclic policy is 106.5, which is fairly similar to the cost of 109.7 incurred in the mixed system in Table I when $f_i = 100$. This small cost discrepancy suggests that the critical due-date lead time threshold for this problem is less than $f_i = 100$. Unfortunately, it is very difficult to calculate the critical threshold due-date lead time for the optimal policy when setup times are present.

An alternative comparison where we can calculate the threshold value is to contrast the mixed system under the hybrid search policy with the SELSP under the generalized base stock policy described in §3.1 of MRW. This generalized base stock policy is similar in form to the hybrid HTAP policy, except it includes a nontrivial idling threshold not present in the hybrid policy. By our analysis in §4.1, the threshold due-date lead time is v_1/ρ_1, where v_1 is the base stock level from the SELSP case. For this SELSP scenario, v_1 was found to be 49.0, making the critical due-date lead time, f_1, equal to 81.6. Hence, heavy traffic theory predicts that the mixed system cost of $f_i = 100$ should be roughly equal to the SELSP cost. Referring again to row 13 of Table VIII of MRW, we find that the SELSP cost under the generalized base stock policy is 117.0. As we predicted, this

value is reasonably close to the mixed system cost of 122.8 in Table I, particularly since part of this discrepancy is probably due to the increased sophistication of the generalized base stock policy relative to the hybrid HTAP policy.

7. CONCLUDING REMARKS

Theories for scheduling a manufacturing facility have examined the issues of customized-standardized product mix, setup penalties and due-dates, but have typically focused on each of them separately. In this paper, Coffman, Puhalskii and Reiman's heavy traffic averaging principle is used to investigate the composite problem. With it, we outline a computational method to optimize within the class of dynamic cyclic policies. For the case where each product has its own deterministic due-date lead time, we qualitatively describe the policy. Our methodology, particularly the determination of how due-dates affect the behavior of the fluid system, may be useful for studying other systems with delay constraints, such as due-date scheduling problems arising in computer applications, inventory-routing problems with time windows, and inventory management of perishable products.

Our heavy traffic analysis suggests that the risks inherent in the uncertainty of random demand and service processes cannot be removed. Yet, by proper scheduling, the impact of the variability can be reduced by channeling the fluctuations in order queues and finished goods inventories into low cost regions of the state space. The presence of setups, due-dates and product mix each dictate how this dampening of cost is performed. Our results yield a simple interpretation of how these facets affect the switching curves that characterize the proposed policy:

due-dates	=	shifts
setups	=	breadth
customized/standardized goods	=	presence/absence of orthant boundaries.

The simplicity of the first of these three observations - that deterministic due-dates merely shift the optimal switching curves - is particularly striking, given how notoriously difficult it is to analyze due-date scheduling problems in a dynamic stochastic setting. This three-point guide allows us to qualitatively understand the nature of dynamic stochastic

scheduling problems with deterministic due-date lead times without explicitly calculating the solution.

Our analysis also sheds light on the value of foreknowledge of customer demand (in the form of due-dates) when the system is heavily loaded, and how this information affects the optimal policy. We find that the costs incurred by standardized products are independent of the due-date lead times for these products. When due-date lead times increase beyond a certain level, the standardized products are made-to-order, but there is no change in cost; orders are filled before their due-date, and these early orders play the same role as a finished goods inventory: They provide a buffer against backordering. Although costs for customized products initially decrease as due-date lead times increase, when the (deterministic) due-date lead times reach a critical value the costs level off and the customized product incurs nearly the same cost as standardized products. In essence, large due-date lead times blur the make-to-order/make-to-stock distinction: They cause standardized products to be made-to-order and provide customized products with the flexibility to be produced early, thereby imitating a make-to-stock mode of production.

ACKNOWLEDGMENT

This research was supported by a grant from the Leaders for Manufacturing Program at MIT and NSF grant DDM-9057297. We thank Marty Reiman for helpful comments and thank Frank Kelly for sharing with us an early draft of Aldous, Kelly and Lehozsky (1995), which deepened our understanding of the problem.

APPENDIX 1: THE MARKOV CHAIN APPROXIMATION ALGORITHM

This supplement describes the Markov chain approximation algorithm that solves equations (23)-(24). The Markov chain is created by discretizing the one dimensional total workload state space into intervals of size h and time into blocks of size Δt^h. If we define $Q^h = \sigma^2 + |ch - s|$ then the transition probabilities of the Markov chain are

$$P^h(w, w-h) = \frac{\sigma^2 + 2h\left(\frac{s}{\tau(w)} - c\right)^+}{2Q^h}. \tag{61}$$

$$P^h(w, w+h) = \frac{\sigma^2 + 2h\left(\frac{s}{\tau(w)} - c\right)^-}{2Q^h} \tag{62}$$

and

$$P^h(w, w) = 1 - \frac{\left(\sigma^2 + h\left|c - \frac{s}{\tau(w)}\right|\right)}{Q^h} . \tag{63}$$

and the time interval itself additionally must be set to

$$\Delta t^h = \frac{h^2}{Q^h} . \tag{64}$$

There is one exception: The reflection boundary is never reached and so the transition probability in the feasible region before w_0 is

$$\tilde{P}^h(w_0 + h, w_0 + h) = 1 - P^h(w_0 - h, w_0 - 2h) . \tag{65}$$

All the other transitions have zero probabilities.

We can now write the dynamic programming optimality equations as

$$V(w) = \sum_y \tilde{P}^h(w, y)V(y) + \left(c(x^{c^*}, \tau, w) - g\right)\Delta t^h . \tag{66}$$

Because the Markov chain is a birth-death process, for a policy τ and w_0 the steady-state distribution and gain can be easily calculated from the transition probabilities. Similarly, the potential function can be recursively calculated by

$$V(w+h) = \frac{g - c(x^{c^*}, \tau(w), w) + (1 - \tilde{P}^h(w, w))V(w) - \tilde{P}^h(w, w+h)V(w+h)}{\tilde{P}^h(w, w-h)}. \tag{67}$$

With $V(w)$ and g determined, an improvement iteration on τ can be achieved by performing the optimization embedded in equation (23) and an improvement on w_0 by finding the threshold that minimizes the gain g. The algorithm terminates when τ and w_0 converge.

APPENDIX 2: AN ALGORITHM FOR $\Theta^*(S)$

This appendix describes an algorithm that constructs $\Theta^*(S)$ by finding an initial set for S equal to zero and then tracks how the set evolves as S increases; for clarity we re-

introduce the notation specifying the dependence of the set Θ^* of non-binding products on S. Our task is complicated by two facts: As S increases the optimal solution can jump in region, and there is no guarantee that when a product becomes binding and leaves $\Theta^*(S)$ it does not re-enter for larger S. We simplify our calculations of $\Theta^*(S)$ by including the type of region the cycle center and cycle length imply in our accounting. Let $\Theta^{*I}(S)$ denote $\Theta^*(S)$ when the cycle center and cycle length satisfy the region I conditions, $\Theta^{*II}(S)$ for the region II conditions and $\Theta^{*III}(S)$ for the third region. The algorithm is based on the following eight observations (proofs are provided in Appendix 3):

1. The cycle center x^{c*} and cycle length τ are continuous functions of the effective setup cost per cycle S.

2. The optimal cycle length τ is monotonically increasing with respect to S.

3. If product i is not in Θ^*, then for $\Theta' = \Theta^* \cup \{i\}$ the cycle center $x^{c\prime}$ and cycle length τ' calculated with Θ' satisfy $x_i^{c\prime} < \tau' \rho_i (1 - \rho_i)/2$.

4. If S' and S'' are such that $S' < S''$ and both their respective optimal cycle lengths and cycle centers satisfy region I (III) conditions, then $\Theta^{*I}(S'') \subset \Theta^{*I}(S')$ ($\Theta^{*III}(S'') \subset \Theta^{*III}(S')$).

5. If S is such that the optimal cycle length and cycle centers imply a shift from region II to region I (III), then $\Theta^{*I}(\lim_{\epsilon \to 0} S + \epsilon) = \Theta^{*II}(\lim_{\epsilon \to 0} S - \epsilon)$ ($\Theta^{*III}(\lim_{\epsilon \to 0} S + \epsilon) = \Theta^{*II}(\lim_{\epsilon \to 0} S - \epsilon)$).

6. If S' and S'' are such that $S' < S''$ and both their respective optimal cycle lengths and cycle centers satisfy region II conditions, then $\Theta^{*II}(S'') \subset \Theta^{*II}(S')$.

7. When a condition 1 binding product i changes to condition 2, it remains binding.

8. If S is such that the optimal cycle length and cycle centers imply a shift from region I or III to region II, then $\Theta^{*II}(\lim_{\epsilon \to 0} S + \epsilon)$ can be calculated by an iterative algorithm.

With these eight observations, the algorithm we suggest is simple. From an initial $\Theta^{*I}(0)$, $\Theta^{*II}(0)$ and $\Theta^{*III}(0)$, we track how each evolves as S is increased. Three types of events can change $\Theta^*(S)$: A shift in region, a non-binding customized product can become binding and a *condition 1* binding product can become *condition 2*. Given equations (10)

and (51), we can calculate the range of S before any one of these events occur.

The initial set of non-binding products is chosen as follows. For an effective setup cost of zero, cycle length τ becomes zero as region II vanishes. If the total workload level w is less than $\sum_{i=1}^{N} \rho_i \bar{f}_i$, then there are few orders remaining in the system and it may be advantageous to make binding some of the cheaper customized products. Let $\{\epsilon_1, \epsilon_2, \ldots, \epsilon_N\}$ be an earliness ordering of the products such that $h_{\epsilon_1} \leq h_{\epsilon_2} \leq \cdots \leq h_{\epsilon_N}$. We find $\Theta^{*I}(0)$ by removing products one-by-one. A product ϵ_j is removed if it is next in the ϵ_i list, is customizable and is insufficient for storing the remaining work, i.e. $\sum_{i=1}^{j} \rho_i \bar{f}_i < \sum_{i=1}^{N} \rho_i f_i - w$. The process stops when either a standardized product is reached in the ϵ_i list or the next product ϵ_j implies $\sum_{i=1}^{j} \rho_i f_i > \sum_{i=1}^{N} \rho_i f_i - w$. If w is greater than $\sum_{i=1}^{N} \rho_i f_i$, then $\Theta^{*I}(0)$ is set to $\{1, \ldots, N\}$. Both $\Theta^{*II}(0)$ and $\Theta^{*III}(0)$ are always set to $\{1, \ldots, N\}$.

Given an initial $\Theta^*(0)$ we can find the range of effective setup costs before the set changes. The set changes when the minimum S such that one of the three events occur: A shift in region, the binding of a customized product or the change in condition of a binding product. For a given S' with non-binding set $\Theta^*(S')$ a region change occurs when

$$S = \begin{cases} \xi_1^{\Theta^{*I}(S')} \left(\dfrac{\sum_{i \in \Theta^{*I}(S')} \rho_i f_i - w}{\Xi_1} \right)^2 - \xi_2^{\Theta^{*I}(S')} \\ \qquad \text{region I to region II shift} \\ \xi_3^{\Theta^{*I}(S')} \left(\dfrac{\alpha_{\theta_h^*(S')} \gamma_2 (\sum_{i \in \Theta^{*II}(S')} \rho_i \bar{f}_i - w)}{-\alpha_{\theta_h^*(S')} \gamma_1 - \alpha_{\theta_h^*(S')} \gamma_2 (\sum_{i \notin \Theta^{*II}(S')} \frac{\rho_i (1-\rho_i)}{2}) + \frac{\rho_{\theta_h^*(S')} (1 - \rho_{\theta_h^*(S')})}{2}} \right)^2 \\ \quad - \xi_4^{\Theta^{*II}(S')} (\sum_{i \in \Theta^{*II}(S')} \rho_i \bar{f}_i - w)^2 - \xi_2^{\Theta^{*II}(S')} \\ \qquad \text{region II to region I shift} \\ \xi_3^{\Theta^{*I}(S')} \left(\dfrac{\alpha_{\theta_h^*(S')} \gamma_2 (\sum_{i \in \Theta^{*II}(S')} \rho_i \bar{f}_i - w)}{-\alpha_{\theta_h^*(S')} \gamma_1 - \alpha_{\theta_h^*(S')} \gamma_2 (\sum_{i \notin \Theta^{*II}(S')} \frac{\rho_i (1-\rho_i)}{2}) - \frac{\rho_{\theta_h^*(S')} (1 - \rho_{\theta_h^*(S')})}{2}} \right)^2 \\ \quad - \xi_4^{\Theta^{*II}(S')} (\sum_{i \in \Theta^{*II}(S')} \rho_i f_i - w)^2 - \xi_2^{\Theta^{*II}(S')} \\ \qquad \text{region II to region III shift} \\ \xi_5^{\Theta^{*III}(S')} \left(\dfrac{\sum_{i \in \Theta^{*III}(S')} \rho_i f_i - w}{\Xi_2} \right)^2 - \xi_2^{\Theta^{*III}(S')} \\ \qquad \text{region III to region II shift} \end{cases} . \tag{68}$$

where

$$\Xi_1 = \sum_{\iota \in \{\Theta^{\bullet I}(S') \setminus \theta_h^\bullet(S')\}} \frac{\rho_\iota(1-\rho_\iota)}{b_\iota + h_\iota}\left[\frac{b_\iota - h_\iota}{2} + h_{\theta^\bullet(S')}\right] + \frac{\rho_{\theta_h^\bullet(S')}(1-\rho_{\theta_h^\bullet(S')})}{2} - \sum_{\iota \notin \Theta^{\bullet I}(S')} \frac{\rho_\iota(1-\rho_\iota)}{2} \tag{69}$$

$$\Xi_2 = \sum_{\iota \in \{\Theta^{\bullet III}(S') \setminus \theta_b^\bullet(S')\}} \frac{\rho_\iota(1-\rho_\iota)}{b_\iota + h_\iota}\left[\frac{b_\iota - h_\iota}{2} - b_{\theta_h^\bullet(S')}\right] - \frac{\rho_{\theta_b^\bullet(S')}(1-\rho_{\theta_b^\bullet(S')})}{2} - \sum_{\iota \notin \Theta^{\bullet III}(S')} \frac{\rho_\iota(1-\rho_\iota)}{2}. \tag{70}$$

A customized product becomes binding when

$$S = \begin{cases} \xi_1^{\Theta^{\bullet I}(S')}\left(\frac{\rho_\iota f_\iota (h_\iota + h_\iota)}{\rho_\iota(1-\rho_\iota)(b_\iota + h_{\theta_h^\bullet})}\right)^2 - \xi_2^{\Theta^{\bullet I}(S')} \\ \qquad\qquad \text{region I. } \iota \neq \theta_h^* \text{ binding} \\ \xi_1^{\Theta^{\bullet I}(S')}\left(\frac{\sum_{\iota \in \{\Theta^{\bullet I}(S') \setminus \theta_h^\bullet(S')\}} \rho_\iota \bar{f}_\iota - w}{\Xi_1}\right)^2 - \xi_2^{\Theta^{\bullet I}(S')} \\ \qquad\qquad \text{region I. } \theta_h^* \text{ binding} \\ \xi_3^{\Theta^{\bullet II}(S')}\left(\frac{\rho_\iota f_\iota - (\sum_{\jmath \in \{\Theta^{\bullet II}(S')} \rho_\jmath \bar{f}_\jmath - w)\alpha_\iota \gamma_2}{\alpha_\iota \gamma_1 + \frac{\rho_\iota(1-\rho_\iota)}{2} + \sum_{\jmath \notin \Theta^{\bullet II}(S')} \frac{\rho_\jmath(1-\rho_\jmath)}{2}}\right)^2 - \xi_2^{\Theta^{\bullet II}(S')} \\ \quad - \xi_4^{\Theta^{\bullet II}(S')}(\sum_{\jmath \in \Theta^{\bullet II}(S')} \rho_\jmath \bar{f}_\jmath - w)^2 \\ \qquad\qquad \text{region II. } \iota \text{ binding} \\ \xi_5^{\Theta^{\bullet III}(S')}\left(\frac{\rho_\iota f_\iota (b_\iota + h_\iota)}{\rho_\iota(1-\rho_\iota)(b_\iota - b_{\theta_b^\bullet})}\right)^2 - \xi_2^{\Theta^{\bullet III}(S')} \\ \qquad\qquad \text{region III. } \iota \neq \theta_b^* \text{ binding} \\ \xi_5^{\Theta^{\bullet III}(S')}\left(\frac{\sum_{\iota \in \{\Theta^{\bullet III}(S') \setminus \theta_b^\bullet(S')\}} \rho_\iota f_\iota - w}{\Xi_2}\right)^2 - \xi_2^{\Theta^{\bullet III}(S')} \\ \qquad\qquad \text{region III. } \theta_b^* \text{ binding} \end{cases}. \tag{71}$$

Lastly, product i changes from *condition 1* to *condition 2* when

$$S = \begin{cases} \xi_1^{\Theta^{*I}(S')} \frac{\bar{f}_i}{1-\rho_i} - \xi_2^{\Theta^{*III}(S')} & \text{region I} \\ \xi_3^{\Theta^{*II}(S')} \frac{\bar{f}_i}{1-\rho_i} - \xi_4^{\Theta^{*II}(S')} (\sum_{j \in \Theta^{*II}(S')} \rho_j \bar{f}_j - w)^2 - \xi_2^{\Theta^{*II}(S')} & \text{region II} \\ \xi_5^{\Theta^{*III}(S')} \frac{\bar{f}_i}{1-\rho_i} - \xi_2^{\Theta^{*III}(S')} & \text{region III} \end{cases} . \quad (72)$$

Thus the current set of $\Theta^*(S')$, $\Theta^{*1}(S')$ and $\Theta^{*2}(S')$ is valid for effective setup cost from S' to the minimum S greater than S' in equations (68), (71), (72). At that point, $\Theta^{*I}(S)$, $\Theta^{*II}(S)$ and $\Theta^{*III}(S)$ are updated according to observations 5 through 8 and the process is repeated. The algorithm ends when either all customized productes are binding or all but one of the customized products are binding and the cheapest is the *condition 3* product in region III with $\frac{1}{\beta\tau}(x_i^c - \frac{\tau\rho_i(1-\rho_i)}{2}) < 0$ (in the other regions, increasing cycle length impies that either the customized product would become binding or that the region would eventually shift).

APPENDIX 3: PROOFS OF EIGHT OBSERVATIONS

In this supplement we prove the eight observations stated in Appendix 2.

Observation 1. For a given w and $V'(w)$, the quantity in equation (23) is continuous with continuous derivatives with respect to the cycle length τ ($\tau > 0$), cycle center x^c and effective setup cost per cycle S. In addition, the boundary conditions are also continuous with continuous derivatives with respect to x^c, τ and S. Thus, the Karesh-Kuhn-Tucker necessary optimality conditions change continuously with S. The only way the optimal x^{c*} and τ^* could be discontinuous with respect to an increase in S is if there were multiple optimal solutions for a given S and w. This is not the case since the objective function is convex with respect to cycle center and cycle length and has the unique optimal solution presented in equations (40) and (51).

Observation 2. This is easily seen from equation (51) and the first observation implying the continuity of τ^* during changes of binding products Θ^* and changes in region.

Observation 3. This follows from the construction of Θ^*.

Observation 4. We show this by inducting on the cheapest products. Consider the

region I case: let θ_h^1 be the first cheapest product to become binding in region I at effective setup cost S^1 with cycle length τ^1. All other products with smaller earliness costs must be binding in *condition 2*. The instant product θ_h^1 becomes binding, the cycle center has reached the positivity boundary and so

$$w - \sum_{i \in \Theta^{*2}(S^1)} \frac{\tau^1 \rho_i (1-\rho_i)}{2} - \sum_{i \in \Theta^{*I}(S^1) \setminus \theta_h^1} \left[\rho_i f_i - \frac{\tau^1 \rho_i (1-\rho_i)}{2} \frac{b_i - h_i + 2h_{\theta_h^1}}{b_i + h_i} \right] - \frac{\tau^1 \rho_{\theta_h^1} (1-\rho_{\theta_h^1})}{2} = 0 . \tag{73}$$

In order for the minimum work experienced over the cycle, $x_i^c - \tau \rho_i (1-\rho_i)/2$, to reach the boundary as τ is increased, its derivative must be negative; that is,

$$- \sum_{i \in \Theta^{*2}(S^1)} \frac{\rho_i (1-\rho_i)}{2} + \sum_{i \in \Theta^{*I}(S^1) \setminus \theta_h^1} \frac{\rho_i (1-\rho_i)}{2} \frac{b_i - h_i + 2h_{\theta_h^1}}{b_i + h_i} - \frac{\rho_{\theta_h^1} (1-\rho_{\theta_h^1})}{2} < 0. \tag{74}$$

As S increases, τ also increases, and hence the cheaper binding *condition 2* products can only re-enter in *condition 2* and will not re-enter in region 1. Additional products might leave $\Theta^{*I}(S^1)$ but since $\frac{b_i - h_i + 2h_{\theta_h^1}}{b_i + h_i} < 1$, equation (74) will become more negative. This implies that equation (73) will remain negative for larger τ, and so θ_h^1 cannot re-enter Θ^{*I}. The i^{th} induction on the cheapest product results in an argument identical to the previous one where equations (73) and (74) are modified by replacing θ_h^1 with θ_h^i and S^1 with S^i.

The other more expensive products that become binding in region I also cannot re-enter Θ^{*I}. Since $h_{\theta_h^*(S'')} > h_{\theta_h^*(S')}$, from the above induction we have

$$\frac{\rho_i (1-\rho_i)}{2} \frac{b_i - h_i + 2 + h_{\theta_h^*(S')}}{b_i + h_i} < \frac{\rho_i (1-\rho_i)}{2} \frac{b_i - h_i + 2 + h_{\theta_h^*(S'')}}{b_i + h_i}. \tag{75}$$

Therefore, the cycle center of the more expensive non-binding products are lower for larger S. Since cycle length τ increases with S, if the cycle center becomes less than $\tau \rho_i (1-\rho_i)/2$, and hence binding, it will remain binding.

The same argument holds for region III.

Observation 5. Since cycle centers and cycle length are continuous in S, this is equivalent to the statement that no binding products in region II become non-binding

after a transition to region I. Binding *condition 1* products must have become binding in region I and remain binding by observation 4 (in region III, a *condition 1* product becoming non-binding would violate the region III definition). *Condition 2* products in region I will have a cycle center of the form $\rho_i \bar{f}_i - \tau\alpha_i \cdot \gamma_1 - (\sum_{j\in\Theta^\bullet} \rho_j \bar{f}_j - w')\alpha_i \cdot \gamma_2$ and so the derivative of $x_i^c - \tau\rho_i(1-\rho_i)/2$ with respect to τ is always negative. This implies that a *condition 2* binding cycle center will continue to push against the orthant boundary.

Observation 6. We prove this by examining $x_i^c - \tau\rho_i(1-\rho_i)/2$, the derivative of the minimum amount of work over the course of a cycle, with respect to τ. Once the derivative is negative, we show that it remains so as further products are removed from $\Theta^{\bullet II}$. Thus when a product becomes binding at $x_i^c - \tau\rho_i(1-\rho_i)/2 = 0$, as τ expands it cannot re-enter: The negative derivative would continue to push the cycle center against the orthant boundary. Since no binding products in $\Theta^{\bullet II}(S)$ can become non-binding in the other regions by observation 5, $\Theta^{\bullet II}(S)$ is a nonincreasing set with S.

The derivative of minimum workload in region II for a non-binding product is

$$\begin{aligned}\frac{\partial}{\partial\tau}\left(x_i^c - \frac{\tau\rho_i(1-\rho_i)}{2}\right) &= -\alpha_i\gamma_1 - \frac{\rho_i(1-\rho_i)}{2} - \frac{\frac{\rho_i(1-\rho_i)}{b_i+h_i}}{\Sigma}\sum_{j\notin\Theta^{\bullet II}(S)}\frac{\rho_j(1-\rho_j)}{2} \\ &= -\frac{\frac{\rho_i(1-\rho_i)}{2(b_i+h_i)}}{\Sigma}\left[\sum_{j\in\Theta^{\bullet II}(S)}(2b_i+h_j-b_j)\frac{\rho_j(1-\rho_j)}{b_j+h_j} + \sum_{j\notin\Theta^{\bullet II}(S)}\rho_j(1-\rho_j)\right]. \quad (76)\end{aligned}$$

where $\Sigma = \sum_{j\in\Theta^{\bullet II}(S)} \rho_j(1-\rho_j)/(b_j+h_j)$. Therefore the derivative is negative if and only if

$$\sum_{j\in\Theta^{\bullet II}(S)}(2b_i+h_j-b_j)\frac{\rho_j(1-\rho_j)}{b_j+h_j} + \sum_{j\notin\Theta^{\bullet II}(S)}\rho_j(1-\rho_j) > 0. \quad (77)$$

Thus, at effective setup cost S', if the derivative of product i is negative and a more expensive tardiness product l leaves to form $\Theta^{\bullet II}(S')$, then equation (77) implies

$$\sum_{j\in\Theta^{\bullet II}(S)\cup\{l\}}(2b_i+h_j-b_j)\frac{\rho_j(1-\rho_j)}{b_j+h_j} + \sum_{j\notin\Theta^{\bullet II}(S)\cup\{l\}}\rho_j(1-\rho_j) > 0. \quad (78)$$

which is

$$\sum_{j\in\Theta^{*II}(S)}(2b_i+h_j-b_j)\frac{\rho_j(1-\rho_j)}{b_j+h_j}+\sum_{j\notin\Theta^{*II}(S)}\rho_j(1-\rho_j)+2\rho_l(1-\rho_l)\frac{b_i-b_l}{b_l+h_l}>0. \quad (79)$$

Since $b_i - b_l < 0$, we have

$$\sum_{j\in\Theta^{*II}(S)}(2b_i+h_j-b_j)\frac{\rho_j(1-\rho_j)}{b_j+h_j}+\sum_{j\notin\Theta^{*II}(S)}\rho_j(1-\rho_j)>0, \quad (80)$$

and so the derivative remains negative. If a cheaper tardiness product l leaves forming $\Theta^{*II}(S)$, the derivative of the cheaper product implies

$$\sum_{j\in\Theta^{*II}(S)\cup\{l\}}(2b_l+h_j-b_j)\frac{\rho_j(1-\rho_j)}{b_j+h_j}+\sum_{j\notin\Theta^{*II}(S)\cup\{l\}}\rho_j(1-\rho_j)>0. \quad (81)$$

The left side of (81) is less than

$$\sum_{j\in\Theta^{*II}(S)}(2b_i+h_j-b_j)\frac{\rho_j(1-\rho_j)}{b_j+h_j}+\sum_{j\notin\Theta^{*II}(S)}\rho_j(1-\rho_j) \quad (82)$$

because their difference is

$$\sum_{j\in\Theta^{*II}(S)}2(b_i-b_l)\frac{\rho_j(1-\rho_j)}{b_j+h_j}. \quad (83)$$

Therefore the derivative of the minimum workload over a cycle for product i remains negative after l becomes binding.

Observation 7. If this change occurs in regions I or III, then product i remains binding because in these regions the quantity $\frac{\partial}{\partial\tau}(x_i^c-\frac{\tau\rho_i(1-\rho_i)}{2})$ is always negative for a *condition 2* product. If the change occurs in region II, a *condition 1* product in region II implies that

$$\sum_{j\in\Theta^{*II}(S)\setminus\{i\}}\frac{\rho_j(1-\rho_j)}{2(b_j+h_j)}[b_j-h_j+2h_i]-\frac{\rho_i(1-\rho_i)}{2}<0. \quad (84)$$

or

$$\rho_i(1-\rho_i)+\sum_{j\in\Theta^{*II}(S)\setminus\{i\}}\frac{\rho_j(1-\rho_j)}{b_j+h_j}[h_j-b_j-2h_i]<0. \quad (85)$$

Subtracting equation (77) from equation (84) we get

$$\sum_{j \in \Theta^{*II}(S) \setminus \{i\}} 2(b_i + h_i) \frac{\rho_j(1-\rho_j)}{b_j + h_j}. \tag{86}$$

which is positive. As stated in observation 6, this implies that $\frac{\partial}{\partial \tau}(x_i^* - \frac{\tau \rho_i(1-\rho_i)}{2})$ is negative.

Observation 8. A conclusion from the previous seven observations is that $\Theta^*(S)$ is relatively predictable with the notable exception of transitions from regions I and III to region II. At these transitions, binding products may again enter $\Theta^*(S)$. Re-calculation of $\Theta^{*II}(S)$, however, is not difficult. At the effective setup cost S point of transition from region I or III to region II, all binding products not before in $\Theta^{*II}(S')$ for $S' < S$ should be re-included in $\Theta^{*II}(S)$. The cycle center can then be re-calculated. Those products such that either their cycle centers are infeasible or were previously binding and currently have a negative cycle center derivative (as determined by equation (77)) can be removed from $\Theta^{*II}(S)$. By observation 6, they do no re-enter. This process can be repeated until $\Theta^{*II}(S)$ is found such that all binding products with negative minimum workload per cycle derivatives are removed. The resulting $\Theta^{*II}(S)$ is equal to $\Theta^{*II}(S^+)$.

REFERENCES

Aldous, D., F. P. Kelly and J. P. Lehoczky. 1995. Working notes. Statistical Laboratory, U. of Cambridge, UK.

Anupindi, R. and S. Tayur. 1994. Managing Stochastic Multi-Product Systems: Model, Measures, and Analysis. Graduate School of Industrial Administration, Carnegie Mellon University, Pittsburgh, PA.

Bertsekas, D. 1995. *Nonlinear Programming*. Athena Scientific, Belmont, MA.

Bertsimas, D. and J. Niño-Mora. 1996. Conservation Laws, Extended Polymatroids and Multiarmed Bandit Problems: A Polyhedral Approach to Indexable Systems. *Math. of Operations Research* **21**, 257-306.

Bertsimas, D. and H. Xu. 1993. Optimization of Polling Systems and Dynamic Vehicle Routing Problems on Networks. Sloan School of Management, MIT, Cambridge, MA.

Boxma, O., H. Levy and J. Westrate. 1994. Efficient Visit Frequencies for Polling Tables: Minimization of Waiting Cost. *Queueing Systems* **9**. 133-162.

Boxma, O. J. and H. Takagi. 1992. Editors, Special Issue on Polling Systems, *Queueing Systems* **11** (1 and 2).

Bourland, K. and C. Yano. 1995. The Strategic Use of Capacity Slack in the Economic Lot Scheduling Problem with Random Demand. *Management Science* **40**. 1690-1704.

Browne, S. and U. Yechiali. 1989. Dynamic Priority Rules for Cyclic-Type Queues. *Advances in Applied Probability* **21**. 432-450.

Buzacott, J. A. and J. G. Shanthikumar. 1993. *Stochastic Models of Manufacturing Systems*. Prentice-Hall, Englewood Cliffs, NJ.

Carr, A., Z. A. Gullu, P. Jackson and J. Muckstadt. 1993. An Exact Analysis of a Production-Inventory Strategy for Industrial Planners. School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY.

Coffman, E. G., Jr., A. A. Puhalskii, and M. I. Reiman. 1995a. Polling Systems with Zero Switchover Times: A Heavy-Traffic Averaging Principle. *Annals of Applied Probability* **5**. 681-719.

Coffman, E. G., Jr., A. A. Puhalskii, and M. I. Reiman. 1995b. Polling Systems with Switchover Times: A Heavy-Traffic Averaging Principle. Bell Laboratories, Lucent Technologies, Murray Hill, NJ.

Cox, D. and W. Smith. 1961. *Queues*. Chapman and Hall, London.

Duenyas, I. and M. P. Van Oyen. 1995. Stochastic Scheduling of Parallel Queues with Set-up Costs. To appear in *QUESTA*.

Duenyas, I. and M. P. Van Oyen. 1996. Heuristic Scheduling of Parallel Heterogeneous Queues with Set-ups. *Management Science* **42**, 814-829.

Elmaghraby, S. E. 1978. The Economic Lot Scheduling Problem (ELSP): Review and Extensions. *Management Science* **24**, 587-598.

Federgruen, A. and Z. Katalan. 1995a. Make-to-Stock or Make-to-Order: That is the Question: Novel Answers to an Ancient Debate. Graduate School of Business, Columbia University, New York, NY.

Federgruen, A. and Z. Katalan. 1995b. Determining Production Schedules under Base-Stock Policies in Single Facility Multi-Item Production Systems. To appear in *Operations Research*.

Federgruen, A. and Z. Katalan. 1996. The Stochastic Economic Lot Scheduling Problem: Cyclic Base-stock Policies with Idle Times. *Management Science* **42**, 783-796.

Gallego, G. 1990. Scheduling the Production of Several Items with Random Demands in a Single Facility. *Management Science* **36**, 1579-1592.

Graves, S. C. 1980. The Multi-Product Production Cycling Problem. *AIIE Transactions* **12**, 233-240.

Ha, A. 1993. Optimal Dynamic Scheduling Policy for a Make-to-Stock Production System. To appear in *Operations Research*.

Hariharan, R. and P. Zipkin. 1995. Customer-order Information, Leadtimes, and Inventories. *Management Science* **41**, 1599-1607.

Hofri, M. and K. W. Ross. 1987. On the Optimal Control of Two Queues with Server Set-up Times and its Analysis. *SIAM Journal of Computing* **16**, 399-420.

Iglehart, D. and W. Whitt. 1970. Multiple Channel Queues in Heavy Traffic, I and II. *Advances in Applied Probability* **2**, 150-177.

Koole, G. 1994. Assigning a Single Server to Inhomogeneous Queues with Switching Costs. Report BS-R9405, CWI, Amsterdam, The Netherlands.

Kushner, H. J. 1977. *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations.* Academic Press, New York, NY.

Kushner, H. J. and P. G. Dupuis. 1992. *Numerical Methods for Stochastic Control Problems in Continuous Time.* Springer-Verlag, New York, NY.

Leachman, R. C. and A. Gascon. 1988. A Heuristic Scheduling Policy for Multi-Item, Single Machine Production Systems with Time-Varying Stochastic Demands. *Management Science* **34**, 377-390.

Liu, Z., P. Nain and D. Towsley. 1992. On Optimal Polling Policies. *Queueing Systems* **11**, 59-83.

Mandl, P. 1968. *Analytical Treatment of One-Dimensional Markov Processes.* Springer-Verlag, New York, NY.

Markowitz, D.M., M. I. Reiman and L. M. Wein. 1995. The Stochastic Economic Lot Scheduling Problem: Heavy Traffic Analysis of Dynamic Cyclic Policies. Sloan School of Management, MIT, Cambridge, MA.

Nguyen, V. 1995a. Fluid and Diffusion Approximations of a Two-station Mixed Queueing Network. *Mathematics of Operations Research* **20**, 321-354.

Nguyen, V. 1995b. On Base-Stock Policies for Make-to-Order/Make-to-Stock Production. Sloan School of Management, MIT, Cambridge, MA.

Pandelis, D. and D. Teneketzis. 1995. Optimal Stochastic Dynamic Scheduling in Multi-Class Queues with Tardiness and/or Earliness Penalties. *Probability Engineering and Information Sciences* **8**, 491-509.

Peña, A. and P. Zipkin. 1993. Dynamic Scheduling Rules for a Multiproduct Make-to-Stock Queue. To appear in *Operations Research.*

Qiu, J. and R. Loulou. 1995. Multiproduct Production/Inventory Control under Random Demands. *IEEE Transactions on Automatic Control* **40**, 350-356.